Matematica della Cognizione Umana: Un'Esplorazione dei Framework Quantitativi della Mente

La Matematizzazione della Mente

La cognizione umana è senza dubbio uno dei fenomeni più complessi dell'universo conosciuto. La capacità della mente di percepire, apprendere, ricordare, ragionare, decidere e comunicare emerge da interazioni intricate tra miliardi di neuroni organizzati in reti altamente strutturate. Per comprendere questi processi in modo rigoroso e predittivo, i ricercatori hanno sviluppato sofisticati framework matematici che catturano aspetti essenziali del funzionamento cognitivo. Questi strumenti matematici vanno dalla teoria dell'informazione alla geometria differenziale, dall'algebra lineare alla teoria dei sistemi dinamici, dalla probabilità bayesiana alla teoria dei giochi.

La matematizzazione della cognizione non è semplicemente un esercizio di formalizzazione astratta. I modelli matematici forniscono predizioni quantitative testabili, rivelano principi computazionali generali, unificano fenomeni apparentemente disparati e guidano lo sviluppo di tecnologie cognitive artificiali. Inoltre, la matematica fornisce un linguaggio preciso e universale che trascende le ambiguità del linguaggio naturale, permettendo comunicazione rigorosa e accumulo progressivo di conoscenza.

Questo saggio esplora in profondità i principali framework matematici utilizzati nello studio della cognizione umana, dalla teoria dell'informazione che quantifica il contenuto informativo degli stimoli sensoriali, ai modelli bayesiani che descrivono l'inferenza probabilistica, dai sistemi dinamici che catturano la temporalità dei processi mentali, alle geometrie non euclidee che rappresentano gli spazi concettuali. L'obiettivo è fornire una comprensione completa e tecnicamente accurata di come la matematica avanzata illumina i meccanismi della mente.

Teoria dell'Informazione e Cognizione

La teoria dell'informazione, sviluppata da Claude Shannon nel 1948, fornisce un framework fondamentale per quantificare e analizzare il processamento dell'informazione nei sistemi cognitivi. Sebbene originariamente concepita per la comunicazione elettronica, i suoi principi si sono rivelati straordinariamente applicabili alla comprensione della cognizione.

Entropia e Contenuto Informativo

Il concetto centrale della teoria dell'informazione è l'entropia, che quantifica l'incertezza o il contenuto informativo di una distribuzione di probabilità. Per una variabile casuale discreta X

con possibili valori x_1, x_2, ..., x_n e distribuzione di probabilità P(X = x_i) = p_i, l'entropia di Shannon H(X) è definita come:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

dove il logaritmo in base 2 misura l'informazione in bit. L'entropia raggiunge il suo massimo quando tutti gli eventi sono equiprobabili, ovvero quando $p_i = 1/n$ per tutti gli i, dando $H(X)_{max} = \log_2(n)$. Al contrario, l'entropia è zero quando un evento è certo, $p_i = 1$ per qualche i.

Nel contesto della cognizione, l'entropia può quantificare l'incertezza percettiva. Per esempio, quando un osservatore cerca di identificare quale di n stimoli è presente in condizioni di rumore, l'entropia prima dell'osservazione cattura l'incertezza iniziale. Dopo l'osservazione, l'informazione acquisita riduce questa incertezza. La riduzione dell'entropia, chiamata informazione mutua, quantifica quanto viene appreso.

Informazione Mutua e Dipendenza Statistica

L'informazione mutua I(X;Y) tra due variabili casuali X e Y quantifica quanto una variabile "dice" sull'altra. È definita come:

$$I(X;Y) = \sum\sum P(x,y) \log_2(P(x,y)/(P(x)P(y)))$$

dove la somma è su tutti i possibili valori di X e Y. L'informazione mutua è simmetrica, I(X;Y) = I(Y;X), ed è sempre non negativa. È zero se e solo se X e Y sono statisticamente indipendenti.

L'informazione mutua può essere espressa anche come la riduzione dell'entropia:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

dove H(X|Y) è l'entropia condizionale di X dato Y, definita come:

$$H(X|Y) = -\sum\sum P(x,y) \log_2(P(x|y))$$

Questa formulazione è particolarmente rilevante per la cognizione: H(X) rappresenta l'incertezza iniziale, H(X|Y) rappresenta l'incertezza residua dopo l'osservazione di Y, e la loro differenza è l'informazione acquisita.

Nel sistema visivo, per esempio, l'informazione mutua tra i pattern di attività neuronale e le caratteristiche dello stimolo visivo quantifica quanta informazione sui stimoli viene codificata dai neuroni. Studi hanno mostrato che singoli neuroni nella corteccia visiva primaria possono trasmettere circa 2-3 bit di informazione per spike su aspetti come l'orientamento delle linee.

Il Canale di Comunicazione Neurale

Il sistema nervoso può essere modellato come un canale di comunicazione dove gli stimoli esterni sono gli input, i pattern di attività neuronale sono gli output, e il rumore neurale introduce incertezza. La capacità del canale C, definita come il massimo di informazione

mutua su tutte le possibili distribuzioni di input, quantifica il limite superiore del tasso di trasmissione dell'informazione:

$C = \max_{P(X)} I(X;Y)$

Per un canale gaussiano con potenza di rumore N e potenza del segnale S, la celebre formula di Shannon-Hartley specifica la capacità:

$C = (1/2) \log_2(1 + S/N)$

dove S/N è il rapporto segnale-rumore. Questa formula ha implicazioni profonde per la cognizione: suggerisce che migliorare la trasmissione dell'informazione richiede aumentare il segnale o ridurre il rumore, e che c'è un limite fondamentale alla velocità di trasmissione determinato dalla fisica del sistema.

Nel sistema uditivo, per esempio, la capacità del canale determina quanta informazione può essere estratta da stimoli acustici complessi. Esperimenti comportamentali hanno mostrato che gli umani possono discriminare circa 7 ± 2 categorie lungo una singola dimensione sensoriale, un risultato che George Miller nel 1956 collegò alle limitazioni della capacità informativa.

Codifica Efficiente e Principi di Parsimonia

Un principio fondamentale derivato dalla teoria dell'informazione è che i sistemi neurali dovrebbero codificare l'informazione in modo efficiente, massimizzando l'informazione trasmessa minimizzando le risorse utilizzate. Questo principio di "efficient coding" o "infomax" è stato proposto da Horace Barlow nel 1961 e ha avuto influenza profonda sulla neuroscienze cognitive.

Il codice ottimale secondo Shannon minimizza la lunghezza media del codice rispettando il vincolo di decodificabilità unica. Per messaggi con probabilità p_i, la lunghezza media ottimale L del codice binario soddisfa:

$H(X) \leq L < H(X) + 1$

Il codice di Huffman realizza questo limite, assegnando codici più corti ai messaggi più probabili. Nel cervello, questo suggerirebbe che stimoli comuni dovrebbero essere rappresentati con pattern di attività neuronale "economici" (pochi spike), mentre stimoli rari possono richiedere pattern più elaborati.

Una variante importante è il "sparse coding", dove gli stimoli vengono rappresentati come combinazioni lineari di un piccolo numero di elementi base da un dizionario sovracompleto. Matematicamente, dato uno stimolo x, si cerca una rappresentazione s tale che:

$x \approx As$

dove A è la matrice del dizionario e s è il vettore di coefficienti, con il vincolo che s sia sparso (la maggior parte dei componenti sono zero). Questo si formalizza come problema di ottimizzazione:

$\min_s (||x - As||^2 + \lambda||s||_0)$

dove $||s||_0$ conta il numero di componenti non zero (norma L0) e λ controlla il trade-off tra fedeltà di ricostruzione e sparsità.

Evidenze sperimentali suggeriscono che i neuroni nella corteccia visiva primaria implementano una forma di sparse coding. I campi recettivi di questi neuroni corrispondono a funzioni base simili a wavelets di Gabor, che emergono naturalmente dall'ottimizzazione per sparse coding di immagini naturali.

Predizione Lineare e Ridondanza

Molti stimoli naturali contengono ridondanza statistica: pixel vicini in un'immagine tendono ad avere valori simili, suoni successivi in un discorso sono correlati. La teoria dell'informazione suggerisce che il cervello dovrebbe rimuovere questa ridondanza per codificare l'informazione in modo efficiente.

La predizione lineare fornisce un framework per quantificare e rimuovere la ridondanza. Dato un segnale temporale $x(t)$, si cerca di predire $x(t)$ come combinazione lineare di valori passati:

$x_{pred}(t) = \sum_{k=1}^{p} a_k x(t-k)$

Il segnale residuo o errore di predizione è:

$e(t) = x(t) - x_{pred}(t)$

Se la predizione è buona, l'errore contiene meno ridondanza del segnale originale. I coefficienti a_k che minimizzano la varianza dell'errore $E[e(t)^2]$ soddisfano le equazioni di Yule-Walker:

$\sum_{k=1}^{p} a_k R(j-k) = -R(j)$ per $j = 1,...,p$

dove $R(k) = E[x(t)x(t-k)]$ è la funzione di autocorrelazione.

Nel contesto della visione, la codifica predittiva suggerisce che i neuroni dovrebbero rappresentare la differenza tra l'input attuale e quello predetto basato su input recenti. Questo è consistente con l'osservazione che i neuroni nella retina e nel nucleo genicolato laterale implementano filtri centro-periferia che enfatizzano i bordi e i contrasti piuttosto che le intensità uniformi.

Teoria dell'Informazione e Apprendimento

La teoria dell'informazione fornisce anche metriche per quantificare l'apprendimento. La divergenza di Kullback-Leibler (KL divergence) misura la differenza tra due distribuzioni di probabilità P e Q:

$D_{KL}(P||Q) = \sum P(x) \log(P(x)/Q(x))$

La divergenza KL è sempre non negativa e è zero se e solo se P = Q. È asimmetrica, D_KL(P||Q) ≠ D_KL(Q||P), e non è una distanza metrica.

Durante l'apprendimento, un organismo aggiorna la sua rappresentazione interna del mondo da una distribuzione prior Q a una distribuzione posterior P basata su nuove evidenze. La divergenza KL D_KL(P||Q) quantifica quanto è cambiata la rappresentazione.

Un'applicazione importante è nell'inferenza variazionale, dove si cerca di approssimare una distribuzione posteriore intratabile P(θ|x) con una distribuzione più semplice Q(θ). Si minimizza la divergenza KL:

min_Q D_KL(Q(θ)||P(θ|x))

Questo è equivalente a massimizzare il "Evidence Lower Bound" (ELBO):

ELBO = E_Q[log P(x|θ)] - D_KL(Q(θ)||P(θ))

che bilancia la verosimiglianza dei dati sotto il modello e la complessità del modello misurata dalla divergenza dal prior.

Probabilità Bayesiana e Inferenza Cognitiva

I modelli bayesiani forniscono un framework matematicamente rigoroso per comprendere come gli organismi inferiscono lo stato del mondo da osservazioni incerte. L'approccio bayesiano alla cognizione, sviluppato intensivamente a partire dagli anni 1990, propone che il cervello implementa inferenza probabilistica ottimale combinando conoscenze a priori con evidenze sensoriali.

Il Teorema di Bayes e l'Inferenza Percettiva

Il teorema di Bayes è il fondamento dell'inferenza probabilistica. Nella sua forma più semplice, dato un'osservazione o evidenza E e un'ipotesi o stato del mondo H, il teorema specifica come aggiornare la probabilità di H:

P(H|E) = P(E|H)P(H)/P(E)

dove P(H) è la probabilità a priori di H prima dell'osservazione, P(E|H) è la verosimiglianza dell'evidenza dato H, P(H|E) è la probabilità a posteriori di H data l'evidenza, e P(E) = ∑_H P(E|H)P(H) è la probabilità marginale dell'evidenza.

Nella percezione, H potrebbe rappresentare lo stato del mondo esterno (per esempio, la posizione di un oggetto) e E rappresenta l'input sensoriale (per esempio, il pattern di attivazione sulla retina). Il teorema di Bayes descrive come il cervello dovrebbe combinare l'informazione sensoriale noisy con aspettative a priori per inferire lo stato del mondo più probabilmente.

Per esempio, considera il problema di stimare la posizione x di un oggetto visivo basandosi su un segnale sensoriale noisy s. Se assumiamo che il rumore sia gaussiano:

$P(s|x) = (1/\sqrt{(2\pi\sigma_s^2)}) \exp(-(s-x)^2/(2\sigma_s^2))$

e che il prior sulla posizione sia anch'esso gaussiano:

$P(x) = (1/\sqrt{(2\pi\sigma_p^2)}) \exp(-(x-\mu_p)^2/(2\sigma_p^2))$

allora il posterior è anche gaussiano:

$P(x|s) \propto \exp(-(x-\mu_{post})^2/(2\sigma_{post}^2))$

dove la media posteriore è una media pesata del prior e della verosimiglianza:

$\mu_{post} = (\sigma_p^2/(\sigma_s^2 + \sigma_p^2))s + (\sigma_s^2/(\sigma_s^2 + \sigma_p^2))\mu_p$

e la varianza posteriore è:

$1/\sigma_{post}^2 = 1/\sigma_s^2 + 1/\sigma_p^2$

Questa formula rivela un principio importante: la stima ottimale è un compromesso tra l'evidenza sensoriale e il prior, con il peso relativo determinato dalle incertezze relative. Quando il rumore sensoriale è alto (σ_s grande), la stima si basa più sul prior. Quando il prior è vago (σ_p grande), la stima si basa più sull'evidenza.

Massima Probabilità A Posteriori e Minimi Quadrati

Una strategia decisionale comune è scegliere l'ipotesi con la massima probabilità a posteriori (MAP):

$x_{MAP} = \text{argmax}_x P(x|s)$

Per i prior e likelihood gaussiani discussi sopra, la stima MAP coincide con la media posteriore μ_{post}. Tuttavia, in generale, MAP e media posteriore possono differire.

Un caso speciale importante è quando il prior è uniforme, $P(x) = \text{const}$. In questo caso:

$x_{MAP} = \text{argmax}_x P(s|x)$

che è la stima di massima verosimiglianza (ML). Con likelihood gaussiana, questo diventa:

$x_{ML} = \text{argmin}_x (s-x)^2$

che è il classico stimatore ai minimi quadrati.

Inferenza in Modelli Gerarchici

Molti problemi cognitivi richiedono inferenza in modelli gerarchici dove variabili latenti a diversi livelli di astrazione devono essere inferite. Per esempio, comprendere una scena visiva richiede inferire simultaneamente le identità degli oggetti, le loro pose, l'illuminazione, e i parametri della camera.

In un modello gerarchico a due livelli, potremmo avere variabili di alto livello θ e variabili di basso livello z che generano le osservazioni x:

$P(x,z,\theta) = P(x|z)P(z|\theta)P(\theta)$

L'inferenza richiede calcolare il posterior congiunto:

$P(z,\theta|x) = P(x|z)P(z|\theta)P(\theta)/P(x)$

Questo è spesso intratabile analiticamente. Un approccio è l'inferenza variazionale che approssima il posterior con una distribuzione fattorizzata:

$Q(z,\theta) = Q(z)Q(\theta)$

e minimizza la divergenza KL tra Q e il vero posterior. Le distribuzioni ottimali soddisfano le equazioni:

$\log Q^*(z) = E_\theta[\log P(x,z,\theta)] + \text{const}$ $\log Q^*(\theta) = E_z[\log P(x,z,\theta)] + \text{const}$

che possono essere risolte iterativamente in un algoritmo chiamato coordinate ascent variational inference.

Filtri di Kalman e Inferenza Dinamica

Molti processi cognitivi sono intrinsecamente temporali: il cervello deve tracciare stati del mondo che cambiano nel tempo basandosi su osservazioni sequenziali. Il filtro di Kalman fornisce una soluzione ottimale per inferenza bayesiana in sistemi lineari gaussiani dinamici.

Il modello assume che lo stato latente x_t evolve secondo:

$x_t = A\, x_{(t-1)} + w_t$

dove A è la matrice di transizione di stato e $w_t \sim N(0,Q)$ è rumore di processo gaussiano. Le osservazioni sono generate da:

$y_t = H\, x_t + v_t$

dove H è la matrice di osservazione e $v_t \sim N(0,R)$ è rumore di osservazione gaussiano.

Il filtro di Kalman mantiene una distribuzione gaussiana sul stato:

$P(x_t|y_{(1:t)}) = N(\mu_t, \Sigma_t)$

e la aggiorna in due fasi. Nella fase di predizione:

$\mu_{t|t-1} = A\, \mu_{(t-1)}$ $\Sigma_{t|t-1} = A\, \Sigma_{(t-1)}\, A^T + Q$

Nella fase di aggiornamento, quando y_t è osservato:

$K_t = \Sigma_{t|t-1}\, H^T (H\, \Sigma_{t|t-1}\, H^T + R)^{-1}$ $\mu_t = \mu_{t|t-1} + K_t (y_t - H\, \mu_{t|t-1})$ $\Sigma_t = (I - K_t\, H)\, \Sigma_{t|t-1}$

dove K_t è chiamato guadagno di Kalman.

Il filtro di Kalman è stato applicato con successo a modellare il tracking visivo, il controllo motorio, e la navigazione. Per esempio, nel controllo motorio, lo stato potrebbe includere la posizione e velocità del braccio, le osservazioni potrebbero includere feedback visivo e propriocettivo, e il filtro di Kalman combina predizioni basate sui comandi motori con feedback sensoriale per stimare lo stato del braccio.

Modelli a Spazio di Stati Non Lineari e Filtri Particellari

Quando le dinamiche o le osservazioni sono non lineari, il filtro di Kalman non è più ottimale. Estensioni come il filtro di Kalman esteso (EKF) linearizzano le funzioni non lineari, mentre il filtro di Kalman unscented (UKF) usa trasformazioni deterministiche per propagare momenti.

Un approccio più generale è il filtro particellare o Monte Carlo sequenziale. Questo rappresenta il posterior con un insieme di particelle (campioni):

$P(x_t|y_{(1:t)}) \approx \sum_{i=1 \text{ to } N} w_t^{(i)} \delta(x_t - x_t^{(i)})$

dove $x_t^{(i)}$ sono le particelle e $w_t^{(i)}$ sono i loro pesi. Le particelle vengono propagate attraverso le dinamiche:

$x_t^{(i)} \sim P(x_t|x_{(t-1)}^{(i)})$

e i pesi vengono aggiornati basandosi sulla verosimiglianza:

$w_t^{(i)} \propto w_{(t-1)}^{(i)} P(y_t|x_t^{(i)})$

Periodicamente, le particelle vengono ricampionate per evitare degenerazione (tutte le particelle tranne una hanno peso trascurabile).

I filtri particellari sono stati usati per modellare processi cognitivi complessi come la categorizzazione di oggetti visivi che cambiano nel tempo o la comprensione del linguaggio in tempo reale.

Campionamento di Gibbs e Monte Carlo a Catena di Markov

Per inferenza in modelli bayesiani complessi, i metodi di Monte Carlo a catena di Markov (MCMC) forniscono un approccio generale. L'idea è costruire una catena di Markov il cui stato stazionario è la distribuzione di interesse.

Il campionamento di Gibbs è una forma di MCMC particolarmente semplice. Per campionare da una distribuzione congiunta $P(x_1,...,x_n)$, si itera ciclicamente campionando ogni variabile condizionatamente a tutte le altre:

$x_1^{(t+1)} \sim P(x_1|x_2^{(t)},...,x_n^{(t)})$ $x_2^{(t+1)} \sim P(x_2|x_1^{(t+1)},x_3^{(t)},...,x_n^{(t)})$ ... $x_n^{(t+1)} \sim P(x_n|x_1^{(t+1)},...,x_{(n-1)}^{(t+1)})$

Sotto condizioni lievi, la distribuzione di $(x_1^{(t)},...,x_n^{(t)})$ converge alla distribuzione target $P(x_1,...,x_n)$.

L'algoritmo Metropolis-Hastings generalizza questo approccio. Dato lo stato corrente x, si propone un nuovo stato x' da una distribuzione di proposta q(x'|x). La proposta è accettata con probabilità:

$\alpha = \min(1, (P(x')q(x|x'))/(P(x)q(x'|x)))$

Questa regola di accettazione garantisce che la catena converga alla distribuzione target.

Recentemente, è stato proposto che il cervello potrebbe implementare inferenza probabilistica attraverso sampling neurale, dove l'attività stocastica dei neuroni rappresenta campioni dalla distribuzione posteriore. Questa prospettiva unifica l'apparente casualità della risposta neuronale con inferenza bayesiana ottimale.

Geometria della Rappresentazione Cognitiva

La cognizione coinvolge la rappresentazione di stimoli, concetti, e conoscenze in spazi astratti. La geometria fornisce strumenti matematici potenti per caratterizzare la struttura di questi spazi rappresentazionali.

Spazi Metrici e Distanze Psicologiche

Un modo fondamentale di organizzare rappresentazioni è attraverso nozioni di similarità o distanza. Matematicamente, uno spazio metrico è un insieme X dotato di una funzione distanza $d: X \times X \to \mathbb{R}$ che soddisfa:

1. Non negatività: $d(x,y) \geq 0$ con uguaglianza se e solo se $x = y$
2. Simmetria: $d(x,y) = d(y,x)$
3. Disuguaglianza triangolare: $d(x,z) \leq d(x,y) + d(y,z)$

Nella cognizione, gli stimoli possono essere rappresentati come punti in uno spazio metrico dove la distanza riflette dissimilarità psicologica. Per esempio, i colori possono essere rappresentati nello spazio CIELAB dove la distanza euclidea approssima differenze percettive.

La distanza euclidea in uno spazio n-dimensionale è:

$d_E(x,y) = \sqrt{(\sum(i=1 \text{ to } n) (x_i - y_i)^2)}$

La distanza di Minkowski generalizza questa:

$d_M(x,y) = (\sum(i=1 \text{ to } n) |x_i - y_i|^p)^{(1/p)}$

dove $p = 1$ dà la distanza di Manhattan (city-block) e $p = 2$ dà la distanza euclidea. Esperimenti di scaling multidimensionale hanno mostrato che p spesso varia con il dominio percettivo: stimoli spettrali (colori, toni) tendono a seguire metriche euclidee ($p = 2$), mentre stimoli definiti da dimensioni separate (forme) tendono a seguire metriche city-block ($p = 1$).

Scaling Multidimensionale

Il scaling multidimensionale (MDS) è una famiglia di tecniche che ricostruisce la struttura geometrica di rappresentazioni da giudizi di similarità. Dato un insieme di stimoli e una matrice di distanze o dissimilarità D dove D_ij quantifica quanto gli stimoli i e j sono dissimili, MDS cerca una configurazione di punti x_1,...,x_n in uno spazio d-dimensionale tale che le distanze d(x_i,x_j) approssimino le dissimilarità D_ij.

Il MDS metrico minimizza una funzione di stress:

Stress = $\sqrt{(\sum_{i<j} (d(x_i,x_j) - D_{ij})^2 / \sum_{i<j} D_{ij}^2)}$

Il MDS non metrico rilassa il vincolo, richiedendo solo che le distanze preservino l'ordine delle dissimilarità:

$d(x_i,x_j) < d(x_k,x_l)$ se e solo se $D_{ij} < D_{kl}$

Il MDS è stato ampiamente utilizzato in psicologia cognitiva per rivelare la struttura di spazi rappresentazionali. Per esempio, è stato applicato a giudizi di similarità di concetti astratti, mostrando che concetti come "libertà", "giustizia", e "democrazia" sono vicini in uno spazio semantico.

Manifolds e Geometria Non Euclidea

Molti spazi rappresentazionali cognitivi non sono ben descritti da geometrie euclidee semplici ma hanno struttura di manifold, ovvero sono spazi che localmente assomigliano a spazi euclidei ma globalmente possono avere topologia complessa.

Un manifold differenziabile di dimensione n è uno spazio topologico M dove ogni punto ha un intorno che può essere mappato diffeomorficamente su $\mathbb{R}^n$. Un esempio semplice è la sfera S^2, che localmente sembra un piano ma globalmente ha curvatura.

Per manifolds Riemanniani, la distanza tra punti è definita geodeticamente: la distanza tra x e y è la lunghezza della curva più corta (geodetica) che li connette. Per un manifold con metrica g, la lunghezza di una curva γ(t) è:

$L[\gamma] = \int_a^b \sqrt{(g_{ij}(\gamma(t)) (d\gamma^i/dt)(d\gamma^j/dt))}\, dt$

dove si usa la convenzione di somma di Einstein sugli indici ripetuti.

La curvatura di un manifold quantifica quanto si discosta dalla geometria euclidea piatta. La curvatura di Gauss K per una superficie bidimensionale può essere calcolata dai coefficienti metrici. Una curvatura positiva (come su una sfera) fa sì che le geodetiche convergano, mentre una curvatura negativa (come su una sella) le fa divergere.

Nel contesto della cognizione, lo spazio di configurazione dei movimenti del corpo umano ha naturalmente struttura di manifold a causa dei vincoli articolari. Le rappresentazioni neurali di questi movimenti potrebbero preservare questa struttura manifold.

Geometria Iperbolica e Gerarchie

Recentemente, è emerso interesse nella geometria iperbolica per rappresentare gerarchie concettuali. Lo spazio iperbolico ha curvatura negativa costante e la proprietà notevole che il numero di punti a distanza r da un punto centrale cresce esponenzialmente con r, permettendo di embedare naturalmente strutture ad albero.

Il modello di Poincaré del disco iperbolico bidimensionale H^2 rappresenta lo spazio iperbolico come il disco unitario {z ∈ ℂ : |z| < 1} con la metrica:

ds^2 = 4(dx^2 + dy^2)/(1 - x^2 - y^2)^2

La distanza iperbolica tra due punti z_1 e z_2 è:

d_H(z_1,z_2) = 2 arctanh(|(z_1 - z_2)/(1 - $\bar{z}$_1 z_2)|)

Le geodetiche nello spazio iperbolico sono archi di cerchio perpendicolari al bordo del disco.

Embeddings iperbolici di concetti linguistici hanno mostrato risultati promettenti, catturando meglio le relazioni gerarchiche rispetto a embeddings euclidei. Per esempio, la gerarchia tassonomica WordNet (cane → mammifero → animale → essere vivente) può essere rappresentata più fedelmente in spazio iperbolico con dimensionalità inferiore.

Topology Algebrica e Persistent Homology

La topology algebrica fornisce strumenti per caratterizzare la forma globale di spazi complessi. L'omologia misura i "buchi" in diverse dimensioni: componenti connesse (H_0), loop (H_1), cavità (H_2), etc.

Per un complesso simpliciale K (una generalizzazione di un grafo che include triangoli, tetraedri, etc.), il k-esimo gruppo di omologia H_k(K) cattura k-cicli che non sono bordi di (k+1)-catene. La dimensione di H_k, chiamata numero di Betti β_k, conta il numero di buchi k-dimensionali.

La persistent homology estende questo analizzando come le caratteristiche topologiche persistono attraverso scale. Dato un insieme di punti in uno spazio metrico, si costruisce una filtrazione di complessi simpliciali a scale crescenti. Per esempio, nel complesso di Čech, si includono tutti i simplessi i cui vertici sono entro distanza ε:

C_ε = {σ : diam(σ) ≤ ε}

Man mano che ε aumenta, nuove caratteristiche topologiche appaiono ("nascono") e scompaiono ("muoiono"). Un barcode o persistence diagram visualizza la vita di queste caratteristiche.

La persistent homology è stata applicata per analizzare la struttura topologica di rappresentazioni neurali. Per esempio, studi hanno rivelato che le popolazioni di place cells nell'ippocampo, che codificano la posizione spaziale, formano una rappresentazione topologica dello spazio fisico: la topologia dello spazio di attivazione neurale preserva la topologia dello spazio fisico.

Teoria dei Grafi e Reti Cognitive

Molte strutture cognitive sono naturalmente rappresentate come grafi: reti semantiche dove i nodi sono concetti e i link sono relazioni, reti di similarità dove i link rappresentano similarità, reti sociali dove i nodi sono persone e i link sono relazioni sociali.

Un grafo G = (V,E) consiste in un insieme di vertici V e un insieme di spigoli E ⊆ V × V. Per grafi pesati, ogni spigolo ha un peso w_ij. La matrice di adiacenza A ha elementi:

A_ij = 1 se (i,j) ∈ E, altrimenti 0

per grafi non pesati, o A_ij = w_ij per grafi pesati.

Metriche importanti sui grafi includono:

Il grado di un nodo i è d_i = ∑_j A_ij, il numero di collegamenti.

La lunghezza del cammino più corto l_ij è il numero minimo di spigoli che connettono i e j.

La lunghezza media del cammino è L = (1/(n(n-1))) ∑_(i≠j) l_ij.

Il coefficiente di clustering di un nodo i quantifica quanto i suoi vicini sono interconnessi:

C_i = (numero di triangoli connessi a i)/(numero di triple connesse a i)

Il coefficiente di clustering medio è C = (1/n) ∑_i C_i.

I grafi small-world, caratterizzati da alto clustering ma breve cammino medio, sono stati proposti come modelli di reti cognitive. Hanno la proprietà che la maggior parte dei nodi può essere raggiunta da ogni altro nodo attraverso pochi passi (piccolo L), ma i nodi tendono a formare cluster locali densi (grande C).

Il modello di Watts-Strogatz genera grafi small-world partendo da un grafo regolare su un anello e "rewirando" casualmente una frazione p di spigoli. Per p piccolo, il grafo mantiene alto clustering ma sviluppa scorciatoie che riducono L.

Analisi di reti semantiche (come free association networks dove le persone producono parole associate a parole cue) hanno rivelato proprietà small-world, suggerendo che la conoscenza semantica è organizzata per permettere accesso rapido ma anche rappresentazione di struttura locale ricca.

Centralità e Importanza nei Grafi

Diverse misure quantificano l'importanza o centralità di nodi in un grafo:

La centralità di grado è semplicemente il grado d_i.

La centralità di vicinanza (closeness) di un nodo i è:

CC_i = (n-1) / ∑_(j≠i) l_ij

inversamente proporzionale alla distanza media da altri nodi.

La centralità di betweenness quantifica quanto spesso un nodo si trova su cammini più corti tra altre coppie:

$BC_i = \sum_{j \neq k \neq i} (\sigma_{jk}(i) / \sigma_{jk})$

dove σ_{jk} è il numero di cammini più corti da j a k e $\sigma_{jk}(i)$ è il numero di questi che passano per i.

La centralità di autovettore assegna importanza a un nodo basandosi sull'importanza dei suoi vicini. È definita dall'equazione di autovettore:

$\lambda x_i = \sum_j A_{ij} x_j$

dove x è l'autovettore associato all'autovalore più grande λ della matrice di adiacenza. L'algoritmo PageRank di Google è una variante di questo.

Nell'analisi di reti semantiche, la centralità può identificare concetti "hub" che sono particolarmente importanti nella struttura concettuale complessiva.

Algebra Lineare e Rappresentazioni Distribuite

L'algebra lineare fornisce il linguaggio matematico per rappresentazioni distribuite dove i concetti sono rappresentati come vettori in spazi ad alta dimensione.

Spazi Vettoriali e Rappresentazioni Semantiche

Nello spazio vettoriale semantico, ogni parola o concetto è rappresentato da un vettore in uno spazio $\mathbb{R}^d$ dove d può essere centinaia o migliaia. La similarità semantica tra due concetti è misurata dalla similarità dei loro vettori, tipicamente usando similarità coseno:

$sim(u,v) = (u \cdot v) / (||u|| \, ||v||) = \sum_i u_i v_i / (\sqrt{\sum_i u_i^2} \sqrt{\sum_i v_i^2})$

La similarità coseno varia tra -1 (vettori opposti) e +1 (vettori identici).

Un metodo classico per costruire rappresentazioni vettoriali è l'analisi semantica latente (LSA). Data una matrice termine-documento X dove X_{ij} è la frequenza della parola i nel documento j, si applica decomposizione a valori singolari (SVD):

$X = U\Sigma V^T$

dove U è una matrice ortogonale n×k (n parole, k dimensioni latenti), Σ è una matrice diagonale k×k con valori singolari, e V è una matrice ortogonale m×k (m documenti). Le righe di U forniscono rappresentazioni vettoriali k-dimensionali delle parole.

La SVD può essere vista come trovare la migliore approssimazione rank-k di X nel senso dei minimi quadrati:

$X_k = U_k \Sigma_k V_k^T = \text{argmin}_{\text{rank}(Y)=k} ||X - Y||_F$

dove $||\cdot||_F$ è la norma di Frobenius.

Word Embeddings e Proprietà Algebriche

Gli embeddings moderni di parole, come Word2Vec e GloVe, hanno rivelato struttura algebrica notevole. Per esempio, relazioni semantiche sembrano corrispondere a trasformazioni lineari:

$\text{vec}(\text{"re"}) - \text{vec}(\text{"uomo"}) \approx \text{vec}(\text{"regina"}) - \text{vec}(\text{"donna"})$

Questa proprietà suggerisce che le relazioni semantiche sono rappresentate come direzioni nello spazio vettoriale.

Il modello Word2Vec Skip-gram definisce la probabilità di una parola di contesto c data una parola target w come:

$P(c|w) = \exp(v_c \cdot v_w) / \sum_{c' \in V} \exp(v_{c'} \cdot v_w)$

dove v_w e v_c sono i vettori di embedding. L'obiettivo è massimizzare la log-likelihood:

$L = \sum_{w \in corpus} \sum_{c \in \text{context}(w)} \log P(c|w)$

GloVe prende un approccio diverso, factorizzando direttamente la matrice di co-occorrenza. L'obiettivo è minimizzare:

$J = \sum_{i,j} f(X_{ij}) (v_i \cdot v_j + b_i + b_j - \log X_{ij})^2$

dove X_{ij} è il numero di volte che le parole i e j co-occorrono, f è una funzione di pesatura, e b_i, b_j sono bias.

Tensori e Rappresentazioni di Ordine Superiore

Mentre i vettori (tensori di ordine 1) e le matrici (tensori di ordine 2) sono sufficienti per molte applicazioni, alcune strutture cognitive richiedono tensori di ordine superiore. Un tensore di ordine n è un array multidimensionale $X \in \mathbb{R}^{(d_1 \times d_2 \times ... \times d_n)}$.

Per esempio, rappresentare conoscenza relazionale può richiedere un tensore di ordine 3 dove X_{ijk} rappresenta la forza della relazione k tra le entità i e j. Il database di knowledge graph potrebbe avere relazioni come "capitale_di(Parigi, Francia)" codificate come $X_{(\text{Parigi,Francia,capitale_di})}$.

La decomposizione tensoriale generalizza la SVD a tensori di ordine superiore. La decomposizione CP (Candecomp/Parafac) approssima un tensore come somma di tensori rank-1:

$X \approx \sum_{r=1 \text{ to } R} \lambda_r a_r \otimes b_r \otimes c_r$

dove ⊗ denota il prodotto esterno. Questo trova applicazione nella modellazione di conoscenza relazionale dove a_r, b_r, c_r rappresentano embeddings latenti di entità e relazioni.

La decomposizione Tucker è più generale:

X ≈ G ×_1 A ×_2 B ×_3 C

dove G è un tensore core e ×_n denota il prodotto tensoriale lungo la n-esima modalità.

Algebre Geometriche e Rappresentazioni Composizionali

Un problema fondamentale è come rappresentare composizioni di concetti. Data rappresentazioni di "rosso" e "automobile", come rappresentare "automobile rossa"? Le semplici operazioni vettoriali (addizione, prodotto) spesso non catturano la struttura composizionale.

Le algebre geometriche forniscono un framework più ricco. In particolare, l'algebra di Grassmann o algebra esterna permette costruire uno spazio di dimensione 2^n da uno spazio base di dimensione n includendo tutti i prodotti esterni (wedge products).

Il prodotto esterno di due vettori u e v è un bivettore u ∧ v che rappresenta il parallelogramma orientato che spannano. Ha le proprietà:

u ∧ v = -v ∧ u (antisimmetria) u ∧ u = 0 (αu) ∧ v = α(u ∧ v) (bilinearità)

Questo permette rappresentare relazioni orientate in modo naturale. Per esempio, "cane morde gatto" potrebbe essere rappresentato come vec(cane) ∧ vec(morde) ∧ vec(gatto), distinto da "gatto morde cane".

Trasformazioni Lineari e Processi Cognitivi

Molti processi cognitivi possono essere modellati come trasformazioni lineari. Una trasformazione lineare T: ℝ^n → ℝ^m soddisfa:

T(αu + βv) = αT(u) + βT(v)

e può essere rappresentata da una matrice M dove T(v) = Mv.

Per esempio, il cambiamento di attenzione può essere modellato come moltiplicazione per una matrice diagonale che amplifica certe dimensioni:

v_attended = Dv

dove D_ii > 1 per dimensioni attese e D_ii < 1 per dimensioni non attese.

La rotazione mentale, dove gli osservatori ruotano mentalmente rappresentazioni di oggetti, può essere modellata come moltiplicazione per matrici di rotazione. In 2D, una rotazione di angolo θ è:

$R(\theta) = [\cos\theta \ -\sin\theta] \ [\sin\theta \ \cos\theta]$

Studi comportamentali hanno mostrato che il tempo di rotazione mentale aumenta linearmente con l'angolo di rotazione, suggerendo un processo analogico di trasformazione continua.

Autovettori e Rappresentazioni Stabili

Gli autovettori di una trasformazione lineare rappresentano direzioni che sono preservate (fino a scala) dalla trasformazione. Per una matrice M, un vettore v è un autovettore con autovalore λ se:

$Mv = \lambda v$

In dinamiche cognitive, gli autovettori possono rappresentare stati stabili o attrattori. Per esempio, in un modello di memoria associativa implementato come una matrice di peso W, i pattern di attivazione che sono autovettori di W sono attrattori: quando il network è inizializzato vicino a un attrattore, converge ad esso.

L'equazione di autovalore è cruciale anche nell'analisi di componenti principali (PCA). Data una matrice di dati X con variabili centrate, la matrice di covarianza è:

$\Sigma = (1/n) X^T X$

Gli autovettori di Σ definiscono le direzioni di varianza massima (componenti principali), e gli autovalori quantificano la varianza lungo queste direzioni. Il primo componente principale v_1 soddisfa:

$v_1 = \text{argmax}_{||v||=1} v^T \Sigma v$

Il PCA è ampiamente usato per riduzione di dimensionalità e visualizzazione di dati cognitivi ad alta dimensione.

Sistemi Dinamici e Processamento Temporale

La cognizione è intrinsecamente temporale. I sistemi dinamici forniscono strumenti matematici per modellare come gli stati cognitivi evolvono nel tempo.

Equazioni Differenziali Ordinarie e Dinamiche Continue

Un sistema dinamico continuo è descritto da equazioni differenziali ordinarie (ODE):

$dx/dt = f(x,t)$

dove $x \in \mathbb{R}^n$ è il vettore di stato e $f: \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ specifica come lo stato cambia. Per sistemi autonomi, f non dipende esplicitamente dal tempo: $dx/dt = f(x)$.

Un esempio semplice ma importante è il modello lineare:

$dx/dt = Ax$

dove A è una matrice n×n. La soluzione è:

$x(t) = \exp(At)\, x(0)$

dove $\exp(At) = \sum_{k=0}^{\infty} (A^k t^k)/k!$ è l'esponenziale di matrice.

Il comportamento qualitativo dipende dagli autovalori di A. Se tutti gli autovalori hanno parte reale negativa, il sistema è stabile e converge a x = 0. Se qualche autovalore ha parte reale positiva, il sistema è instabile.

Nel contesto cognitivo, i modelli di decisione come il drift-diffusion model sono formulati come ODE stocastiche:

$dx = \mu\, dt + \sigma\, dW$

dove μ è il drift (accumulo di evidenza), σ è il coefficiente di diffusione, e dW è un processo di Wiener (moto browniano). L'evidenza x accumula fino a raggiungere una soglia decisionale.

Attrattori e Stabilità

Un concetto centrale è quello di attrattore: un insieme A nello spazio di stato tale che traiettorie che partono vicino ad A convergono ad A. Formalmente, A è un attrattore se:

1. A è invariante: se $x(0) \in A$, allora $x(t) \in A$ per tutto t
2. A è stabile: traiettorie che partono vicino ad A rimangono vicine
3. A ha un bacino di attrazione: esiste un intorno U di A tale che traiettorie da U convergono ad A

I tipi principali di attrattori sono:

Punti fissi: $dx/dt = 0$, il sistema rimane costante

Cicli limite: traiettorie chiuse periodiche

Attrattori strani: attrattori frattali associati a caos

Nella memoria associativa, pattern memorizzati possono essere rappresentati come attrattori. Iniziando da un pattern parziale o degradato (nel bacino), il sistema converge al pattern completo (l'attrattore).

Linearizzazione e Analisi Locale

Per sistemi non lineari $dx/dt = f(x)$, il comportamento vicino a un punto fisso x^* (dove $f(x^*) = 0$) può essere analizzato tramite linearizzazione. Espandendo in serie di Taylor:

$f(x^* + \xi) \approx f(x^*) + Df(x^*)\xi = Df(x^*)\xi$

dove Df è la matrice Jacobiana:

$Df_{ij} = \partial f_i / \partial x_j$

Il comportamento locale è determinato dagli autovalori di Df(x*). Se tutti hanno parte reale negativa, x* è stabile (nodo stabile). Se qualcuno ha parte reale positiva, x* è instabile. Se gli autovalori sono complessi coniugati, il sistema oscilla.

Biforcazioni e Transizioni di Fase Cognitive

Le biforcazioni descrivono cambiamenti qualitativi nel comportamento di un sistema quando un parametro varia. Le biforcazioni principali includono:

Biforcazione saddle-node: due punti fissi (uno stabile, uno instabile) collidono e scompaiono

Biforcazione transcritica: due punti fissi si scambiano la stabilità

Biforcazione pitchfork: un punto fisso diventa instabile e due nuovi punti fissi stabili emergono

Biforcazione di Hopf: un punto fisso stabile diventa instabile e un ciclo limite stabile emerge

Le biforcazioni sono state usate per modellare transizioni qualitative nella cognizione. Per esempio, la transizione da percezione binoculare di due immagini separate a fusione binoculare può essere modellata come biforcazione pitchfork. Le transizioni tra stati percettivi in stimoli ambigui (come il cubo di Necker) possono essere modellate come biforcazioni indotte da rumore.

Caos e Imprevedibilità

I sistemi caotici sono deterministici ma esibiscono sensibilità estrema alle condizioni iniziali: traiettorie che partono arbitrariamente vicine divergono esponenzialmente. La divergenza è quantificata dall'esponente di Lyapunov λ:

$\lambda = \lim_{t\to\infty} (1/t) \log(||\delta x(t)||/||\delta x(0)||)$

Se $\lambda > 0$, il sistema è caotico. Il celebre sistema di Lorenz:

$dx/dt = \sigma(y - x)$ $dy/dt = x(\rho - z) - y$
$dz/dt = xy - \beta z$

è caotico per certi valori dei parametri σ, ρ, β.

È stato proposto che alcune dinamiche neurali siano caotiche, conferendo flessibilità e ricchezza rappresentazionale. Tuttavia, la questione di quanto caos ci sia realmente nelle dinamiche cognitive rimane aperta e dibattuta.

Campi Dinamici Neurali

Per modellare popolazioni di neuroni distribuiti spazialmente, si usano campi dinamici neurali (neural fields). Questi descrivono l'attività u(x,t) in funzione della posizione x e tempo t:

$\partial u(x,t)/\partial t = -u(x,t) + h + \int w(x,x') f(u(x',t)) dx'$

dove h è input esterno, w(x,x') è la funzione di connettività (quanto i neuroni in x influenzano quelli in x'), e f è una funzione di attivazione non lineare.

Per connettività locale eccitatoria e laterale inibitoria (Mexican hat):

$w(x,x') = A_e \exp(-|x-x'|^2/(2\sigma_e^2)) - A_i \exp(-|x-x'|^2/(2\sigma_i^2))$

questo modello può produrre "bumps" di attività localizzati che rappresentano memoria di lavoro di locazioni spaziali.

Oscillatori Accoppiati e Sincronizzazione

Molti processi cognitivi coinvolgono oscillazioni: ritmi theta nell'ippocampo, oscillazioni gamma nella corteccia. Il coupling tra oscillatori può produrre sincronizzazione.

Il modello di Kuramoto descrive N oscillatori con fasi θ_i:

$d\theta_i/dt = \omega_i + (K/N) \sum_{j=1 \text{ to } N} \sin(\theta_j - \theta_i)$

dove ω_i è la frequenza naturale dell'oscillatore i e K è la forza di coupling. Per K sufficientemente grande, gli oscillatori si sincronizzano: $|\theta_i - \theta_j| \to 0$.

Il parametro d'ordine r che misura il grado di sincronizzazione è definito da:

$r e^{i\psi} = (1/N) \sum_{j=1 \text{ to } N} e^{i\theta_j}$

dove r = 1 indica sincronizzazione perfetta e $r \approx 0$ indica fasi casuali.

La sincronizzazione neurale è stata proposta come meccanismo per binding percettivo: caratteristiche di un oggetto (colore, forma, movimento) elaborate in regioni cerebrali diverse vengono "legate insieme" attraverso sincronizzazione delle oscillazioni.

Teoria della Decisione e Scelta Razionale

La teoria della decisione fornisce framework matematici per comprendere come gli organismi fanno scelte in condizioni di incertezza.

Teoria dell'Utilità Attesa

La teoria classica della decisione razionale è la teoria dell'utilità attesa (EUT) di von Neumann e Morgenstern. Dato un insieme di azioni A e stati del mondo S con probabilità P(s), un decisore ha una funzione di utilità $u: A \times S \to \mathbb{R}$ che assegna utilità a ogni combinazione di azione e stato.

L'utilità attesa di un'azione a è:

$EU(a) = \sum_{s \in S} P(s)\, u(a,s)$

Il principio della massimizzazione dell'utilità attesa dice che un decisore razionale sceglie l'azione che massimizza EU:

$a^* = \operatorname{argmax}_{a \in A} EU(a)$

Von Neumann e Morgenstern hanno dimostrato che se le preferenze di un decisore soddisfano certi assiomi (completezza, transitività, continuità, indipendenza), allora esistono probabilità e utilità tali che le preferenze sono rappresentate da massimizzazione dell'utilità attesa.

Avversione al Rischio e Funzioni di Utilità

La forma della funzione di utilità riflette attitudini verso il rischio. Se u è concava ($u'' < 0$), il decisore è avverso al rischio: preferisce un guadagno certo alla sua attesa. Se u è convessa, il decisore è propenso al rischio.

La misura di Arrow-Pratt di avversione al rischio assoluta è:

$r_A(x) = -u''(x)/u'(x)$

Un esempio comune è la funzione di utilità a potenza:

$u(x) = x^{(1-\gamma)}/(1-\gamma)$

dove $\gamma > 0$ è il coefficiente di avversione al rischio relativa. Per $\gamma = 1$, questo diventa $u(x) = \log(x)$.

Esperimenti comportamentali hanno rivelato deviazioni sistematiche dall'EUT. Il celebre paradosso di Allais mostra che le persone violano l'assioma di indipendenza. Il paradosso di Ellsberg mostra avversione all'ambiguità (preferenza per rischi noti su rischi sconosciuti).

Teoria del Prospetto

Kahneman e Tversky hanno proposto la teoria del prospetto come alternativa descrittiva all'EUT che cattura meglio le scelte umane reali. La teoria introduce diverse modifiche:

1. I risultati sono codificati come guadagni o perdite relativi a un punto di riferimento, non come stati finali assoluti

2. La funzione di valore v ha forma a S: concava per guadagni (avversione al rischio), convessa per perdite (propensione al rischio), e più ripida per perdite che per guadagni (avversione alla perdita):

$v(x) = x^{\alpha}$ se $x \geq 0$ $v(x) = -\lambda(-x)^{\beta}$ se $x < 0$

tipicamente con $\alpha,\beta \approx 0{,}88$ e $\lambda \approx 2{,}25$.

3. Le probabilità sono trasformate da una funzione di pesatura w(p) che sovrastima probabilità basse e sottostima probabilità alte:

$$w(p) = p^\gamma / (p^\gamma + (1-p)^\gamma)^{(1/\gamma)}$$

con $\gamma \approx 0{,}61$.

Il valore di un prospetto $(x_1,p_1; x_2,p_2; ...; x_n,p_n)$ è:

$$V = \sum_i w(p_i)\, v(x_i)$$

Teoria dei Giochi e Interazione Strategica

Quando le decisioni coinvolgono interazioni con altri agenti, la teoria dei giochi fornisce il framework appropriato. Un gioco è specificato da:

- Un insieme di giocatori $N = \{1,...,n\}$
- Per ogni giocatore i, un insieme di strategie S_i
- Per ogni giocatore i, una funzione di payoff $u_i: S_1 \times ... \times S_n \to \mathbb{R}$

Un equilibrio di Nash è un profilo di strategie $(s_1^*,...,s_n^*)$ tale che nessun giocatore può migliorare unilateralmente:

$u_i(s_i^*,s_{(-i)}) \geq u_i(s_i,s_{(-i)})$ per tutti $s_i \in S_i$

dove $s_{(-i)}^*$ denota le strategie di tutti i giocatori tranne i.

Per giochi con strategie miste (dove i giocatori randomizzano), un equilibrio di Nash in strategie miste è un profilo di distribuzioni di probabilità $(\sigma_1^*,...,\sigma_n^*)$ tale che ogni strategia nella support di σ_i^* è una best response a $\sigma_{(-i)}^*$.

Il celebre teorema di Nash garantisce l'esistenza di almeno un equilibrio di Nash in strategie miste per ogni gioco finito.

Apprendimento per Rinforzo

L'apprendimento per rinforzo (reinforcement learning, RL) modella come gli agenti apprendono a massimizzare ricompensa cumulativa attraverso interazione con l'ambiente. Il framework standard è il processo decisionale di Markov (MDP).

Un MDP è una tupla (S, A, T, R, γ) dove:

- S è l'insieme degli stati
- A è l'insieme delle azioni
- $T(s'|s,a)$ è la probabilità di transizione allo stato s' da s con azione a
- $R(s,a)$ è la ricompensa immediata
- $\gamma \in [0,1]$ è il fattore di sconto

Una policy π: S → A specifica quale azione prendere in ogni stato. Il valore di uno stato s sotto policy π è:

$V^{\pi}(s) = E_{\pi}[\sum_{t=0}^{\infty} \gamma^t R(s_t,a_t) \mid s_0 = s]$

La policy ottimale π* massimizza il valore:

$V^*(s) = \max_{\pi} V^{\pi}(s)$

La funzione di valore ottimale soddisfa l'equazione di Bellman:

$V^*(s) = \max_a [R(s,a) + \gamma \sum_{s'} T(s'|s,a) V^*(s')]$

Algoritmi come Q-learning apprendono la funzione di azione-valore:

$Q(s,a) = R(s,a) + \gamma \sum_{s'} T(s'|s,a) \max_{a'} Q(s',a')$

attraverso aggiornamenti temporali-differenza:

$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$

dove α è il tasso di apprendimento e r è la ricompensa osservata.

Nel cervello, i neuroni dopaminergici sembrano codificare un segnale di errore di predizione della ricompensa:

$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$

che è precisamente il segnale TD usato nel RL. Questa corrispondenza suggerisce che il cervello implementa algoritmi di RL.

Banditi Multi-Braccio e Exploration-Exploitation

Il dilemma banditi multi-braccio (multi-armed bandit) cattura il trade-off tra exploration (provare azioni per apprendere i loro valori) ed exploitation (scegliere azioni note per essere buone).

Data K azioni con ricompense medie sconosciute $\mu_1,...,\mu_K$, l'obiettivo è massimizzare la ricompensa cumulativa totale. Il regret è la differenza tra la ricompensa ottenuta e quella che si sarebbe ottenuta sempre scegliendo l'azione ottimale:

$\text{Regret}_T = T \mu^* - \sum_{t=1}^{T} r_t$

dove $\mu^* = \max_k \mu_k$.

La strategia ε-greedy seleziona l'azione con il valore stimato più alto con probabilità 1-ε e un'azione casuale con probabilità ε.

La strategia UCB (Upper Confidence Bound) seleziona l'azione che massimizza:

$a_t = \text{argmax}_k [Q_t(k) + c\sqrt{(\log t / N_t(k))}]$

dove Q_t(k) è la stima del valore dell'azione k al tempo t, N_t(k) è il numero di volte che k è stata selezionata, e c controlla exploration.

Il termine √(log t / N_t(k)) fornisce un "bonus di ottimismo": azioni selezionate raramente ricevono bonus più alti, incoraggiando exploration.

La strategia Thompson sampling è bayesiana: mantiene una distribuzione posteriore su μ_k per ogni azione, campiona da queste distribuzioni, e seleziona l'azione con il campione più alto. Questa strategia equilibra naturalmente exploration ed exploitation.

Studi comportamentali suggeriscono che gli umani usano strategie che bilanciano exploration ed exploitation, sebbene spesso con bias verso over-exploitation o exploration diretta più di quanto sarebbe ottimale.

Statistica e Analisi dei Dati Cognitivi

La modellazione quantitativa della cognizione richiede metodi statistici sofisticati per analizzare dati comportamentali e neurali.

Modelli Lineari Generalizzati

I modelli lineari classici assumono:

$y = X\beta + \varepsilon$

dove y è il vettore di osservazioni, X è la matrice di design, β sono i parametri, e $\varepsilon \sim N(0,\sigma^2 I)$ è rumore gaussiano.

I modelli lineari generalizzati (GLM) estendono questo permettendo distribuzioni non gaussiane e relazioni non lineari tra predittori e risposta. Un GLM consiste di:

1. Una distribuzione dalla famiglia esponenziale: $P(y|\theta) = \exp((y\theta - b(\theta))/a(\varphi) + c(y,\varphi))$
2. Un predittore lineare: $\eta = X\beta$
3. Una funzione di link g che connette la media $\mu = E[y]$ al predittore: $g(\mu) = \eta$

Per dati binari (successo/fallimento), si usa regressione logistica con distribuzione binomiale e link logit:

$g(\mu) = \log(\mu/(1-\mu))$

quindi:

$P(y=1|x) = 1/(1 + \exp(-x^T \beta))$

Per dati di conteggio, si usa regressione di Poisson con link log:

$g(\mu) = \log(\mu)$

quindi:

$P(y|x) = \exp(-\lambda(x)) \lambda(x)^y / y!$

dove $\lambda(x) = \exp(x^T \beta)$.

I GLM sono ampiamente usati nell'analisi di dati cognitivi e neurali. Per esempio, le risposte neurali (spike counts) sono spesso modellate con GLM di Poisson.

Modelli Gerarchici e Effetti Misti

I dati cognitivi hanno spesso struttura gerarchica: prove nidificate dentro soggetti, soggetti nidificati dentro gruppi. I modelli gerarchici o modelli ad effetti misti catturano questa struttura.

Un modello lineare misto semplice è:

$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})x_{ij} + \varepsilon_{ij}$

dove y_{ij} è la risposta per il soggetto i nella prova j, β_0 e β_1 sono effetti fissi (parametri di popolazione), b_{0i} e b_{1i} sono effetti casuali (deviazioni specifiche del soggetto), e ε_{ij} è errore residuo.

Tipicamente si assume:

$[b_{0i}] \sim N([0], [\sigma_0^2 \sigma_{01}]) [b_{1i}] [0] [\sigma_{01} \sigma_1^2]$

Gli effetti casuali catturano variabilità tra soggetti, permettendo inferenza sulla popolazione piuttosto che solo sui soggetti osservati.

La stima è tipicamente fatta via massima verosimiglianza ristretta (REML) o metodi bayesiani.

Modelli Gerarchici Bayesiani

L'approccio bayesiano ai modelli gerarchici specifica una gerarchia di distribuzioni:

Livello dati: $y_{ij} \sim P(y|\theta_i)$ Livello soggetto: $\theta_i \sim P(\theta|\varphi)$ Livello popolazione: $\varphi \sim P(\varphi)$

L'inferenza usa il teorema di Bayes:

$P(\theta,\varphi|y) \propto P(y|\theta)P(\theta|\varphi)P(\varphi)$

Questo framework è particolarmente potente per "borrowing strength": le stime per soggetti individuali sono "shrunk" verso la media di gruppo, migliorando stabilità particolarmente per soggetti con pochi dati.

Per esempio, nel modellare tempi di reazione, potremmo avere:

RT_ij ~ LogNormal(μ_i, σ) μ_i ~ N(μ_group, σ_group) μ_group ~ N(500, 100) σ_group ~ HalfCauchy(0, 50)

Selezione di Modelli e Confronto

Con molteplici modelli candidati, serve un modo principiato per selezionare il migliore. Il likelihood ratio test confronta modelli nidificati:

LR = -2(log L_0 - log L_1)

dove L_0 è la likelihood del modello più semplice e L_1 quella del modello più complesso. Sotto l'ipotesi nulla che il modello semplice sia corretto, LR ~ χ^2_k dove k è la differenza nel numero di parametri.

Per modelli non nidificati, si usano criteri di informazione che bilanciano fit e complessità. Il criterio di informazione di Akaike (AIC) è:

AIC = -2 log L + 2k

dove k è il numero di parametri. Si preferisce il modello con AIC più basso.

Il criterio di informazione bayesiano (BIC) usa una penalità più forte:

BIC = -2 log L + k log n

dove n è il numero di osservazioni.

Per modelli bayesiani, si usa il fattore di Bayes:

BF_12 = P(y|M_1) / P(y|M_2)

dove P(y|M) = ∫ P(y|θ,M)P(θ|M)dθ è l'evidenza marginale. Il fattore di Bayes quantifica quanto i dati favoriscono M_1 su M_2.

Analisi di Dati Neurali ad Alta Dimensione

I dati di neuroimaging (fMRI, EEG, MEG) e di registrazioni neurali multi-elettrodo sono intrinsecamente ad alta dimensione. Metodi specializzati sono necessari.

L'analisi di componenti indipendenti (ICA) cerca di decomporre segnali misti in componenti statisticamente indipendenti. Data matrice di osservazione X = AS dove A è la mixing matrix e S contiene sorgenti indipendenti, ICA stima A e S massimizzando la non-gaussianità o minimizzando informazione mutua tra componenti.

Un algoritmo popolare è FastICA che massimizza la neghentropia:

J(y) = H(y_gauss) - H(y)

dove y_gauss è una variabile gaussiana con la stessa varianza di y e H è l'entropia.

Per fMRI, ICA identifica reti funzionali spazialmente distribuite. Per EEG, separa artefatti (movimenti oculari, muscolari) da segnali cerebrali.

La decodifica neurale usa pattern di attività neurale per predire variabili comportamentali o stimoli. Un approccio standard è classificazione con support vector machines (SVM).

Data dati di training $\{(x_i,y_i)\}$ dove x_i è il pattern di attività neurale e $y_i \in \{-1,+1\}$ è la label, SVM trova l'iperpiano separatore che massimizza il margine:

$$\min_{w,b} (1/2)||w||^2$$

soggetto a: $y_i(w^T x_i + b) \geq 1$

Per dati non linearmente separabili, si introduce slack variables ξ_i e si minimizza:

$$(1/2)||w||^2 + C\sum_i \xi_i$$

Per problemi non lineari, il kernel trick permette lavorare in spazi feature ad alta dimensione implicitamente. Il kernel gaussiano è:

$$K(x,x') = \exp(-||x-x'||^2/(2\sigma^2))$$

L'accuratezza di decodifica quantifica quanta informazione sul task è contenuta nei pattern neurali.

Matematica del Linguaggio e Comunicazione

Il linguaggio è un sistema cognitivo complesso che ha attratto considerevole attenzione matematica.

Teoria dell'Informazione del Linguaggio

Shannon applicò la teoria dell'informazione al linguaggio, definendo l'entropia di una lingua come l'incertezza media per simbolo:

$$H = \lim_{n\to\infty} -(1/n) \sum_s P(s) \log P(s)$$

dove la somma è su tutte le sequenze possibili s di n simboli.

Shannon stimò l'entropia dell'inglese usando esperimenti di predizione umana, ottenendo circa 1,0-1,5 bit per lettera. Questo è molto inferiore al $\log_2(26) \approx 4,7$ bit se le lettere fossero equiprobabili, riflettendo la grande ridondanza del linguaggio naturale.

La legge di Zipf descrive la distribuzione di frequenza delle parole: la frequenza f della parola di rango r è approssimativamente:

$$f(r) \propto 1/r^\alpha$$

dove $\alpha \approx 1$. Quindi, la parola più comune è circa due volte più frequente della seconda più comune, tre volte più frequente della terza, etc.

Questa distribuzione può essere derivata da principi di ottimizzazione comunicativa: bilanciare il costo di produzione (favorisce parole corte per concetti frequenti) con precisione comunicativa.

Modelli Formali del Linguaggio

La gerarchia di Chomsky classifica le grammatiche formali per potere espressivo:

Tipo 0 (ricorsivamente enumerabili): generati da macchine di Turing Tipo 1 (context-sensitive): $S \to \alpha A \beta$ implica $\alpha A \beta \to \alpha \gamma \beta$ Tipo 2 (context-free): $A \to \gamma$ Tipo 3 (regolari): $A \to aB$ o $A \to a$

Le grammatiche context-free (CFG) sono particolarmente rilevanti per il linguaggio naturale. Una CFG consiste di:

- Un insieme di simboli non terminali N
- Un insieme di simboli terminali Σ
- Un simbolo iniziale $S \in N$
- Un insieme di regole di produzione P della forma $A \to \alpha$ dove $A \in N$ e $\alpha \in (N \cup \Sigma)^*$

Per esempio, una CFG semplice per frasi inglesi:

$S \to NP\ VP$ $NP \to Det\ N \mid Det\ Adj\ N$ $VP \to V\ NP$ $Det \to \{the, a\}$ $N \to \{cat, dog\}$ $Adj \to \{big, small\}$ $V \to \{chased, saw\}$

Il parsing determina se una stringa è generata dalla grammatica e la sua struttura. L'algoritmo CKY (Cocke-Kasami-Younger) fa parsing in tempo $O(n^3)$ per CFG in forma normale di Chomsky.

Grammatiche Probabilistiche

Le grammatiche probabilistiche context-free (PCFG) assegnano probabilità alle regole:

$P(A \to \alpha)$

con $\sum_\alpha P(A \to \alpha) = 1$ per ogni non terminale A. La probabilità di un albero parse T è il prodotto delle probabilità delle regole usate:

$P(T) = \prod_{r \in T} P(r)$

La probabilità di una stringa w è la somma su tutti i parse:

$P(w) = \sum_T P(T)$

L'inside algorithm calcola P(w) efficientemente usando programmazione dinamica.

Le PCFG permettono modellare ambiguità: stringhe con molteplici parse hanno probabilità distribuita tra di essi.

Modelli Neuronali del Linguaggio

I modelli neuronali recenti hanno rivoluzionato il processamento del linguaggio naturale. I recurrent neural networks (RNN) processano sequenze ricorrentemente:

h_t = f(W_h h_(t-1) + W_x x_t + b)

dove h_t è lo stato nascosto al tempo t, x_t è l'input, e f è una non linearità.

Le Long Short-Term Memory (LSTM) networks introducono porte per controllare il flusso di informazione:

f_t = σ(W_f [h_(t-1), x_t] + b_f) (forget gate) i_t = σ(W_i [h_(t-1), x_t] + b_i) (input gate) C̃_t = tanh(W_C [h_(t-1), x_t] + b_C) (candidate cell state) C_t = f_t ⊙ C_(t-1) + i_t ⊙ C̃_t (cell state) o_t = σ(W_o [h_(t-1), x_t] + b_o) (output gate) h_t = o_t ⊙ tanh(C_t) (hidden state)

dove σ è la funzione sigmoide e ⊙ è prodotto element-wise.

I Transformers, introdotti nel 2017, usano meccanismi di attenzione invece di ricorrenza. Il self-attention calcola:

Attention(Q,K,V) = softmax(QK^T/√d_k)V

dove Q (query), K (key), e V (value) sono proiezioni lineari dell'input, e d_k è la dimensione.

Il multi-head attention usa molteplici "teste" di attenzione in parallelo:

MultiHead(Q,K,V) = Concat(head_1,...,head_h)W^O

dove head_i = Attention(QW_i^Q, KW_i^K, VW_i^V).

I modelli Transformer come GPT e BERT hanno raggiunto performance straordinarie su una vasta gamma di compiti linguistici, sollevando questioni su quanto catturino aspetti della cognizione linguistica umana.

Semantica Distribuzionale

L'ipotesi distribuzionale afferma che parole con significato simile appaiono in contesti simili. Questo motiva rappresentazioni vettoriali dove la similarità semantica corrisponde a vicinanza geometrica.

Il modello Skip-gram with Negative Sampling (SGNS) massimizza:

L = ∑*(w,c)∈D log σ(v_w · v_c)* + ∑(w,c')∈D' log σ(-v_w · v_c')

dove D sono coppie parola-contesto osservate, D' sono coppie negative campionate, σ è la sigmoide, e v_w, v_c sono vettori di word embedding.

GloVe factorizza la matrice di co-occorrenza minimizzando:

J = ∑_(i,j) f(X_ij)(v_i^T v_j + b_i + b_j - log X_ij)^2

dove X_ij è il numero di co-occorrenze e f è una funzione di pesatura.

Gli embeddings risultanti catturano relazioni analogiche: vec("re") - vec("uomo") + vec("donna") ≈ vec("regina").

Analisi Sintattica Computazionale

Il parsing statistico combina grammatiche formali con modelli probabilistici. Il modello generativo di Collins per parsing usa:

P(T|S) ∝ ∏_(r∈T) P_r(rule(r)|parent(r),head(parent(r)))

dove ogni regola è condizionata sul genitore e sulla sua testa lessicale.

I parser neurali moderni usano reti neurali per punteggiare alberi. Il parser transition-based mantiene una configurazione (stack, buffer, archi) e apprende una funzione che predice la prossima azione:

a_t = argmax_a θ^T φ(s_t, a)

dove φ estrae features dalla configurazione s_t.

Recentemente, i parser basati su Transformer processano l'intera sequenza in parallelo, raggiungendo accuratezza superiore al 95% su Penn Treebank.

Reti Neurali Artificiali e Modelli Connessionisti

Le reti neurali artificiali forniscono un ponte tra matematica, neuroscienze e cognizione.

Percettroni e Funzioni Lineari

Il percettrone, introdotto da Rosenblatt nel 1958, è il modello di neurone artificiale più semplice:

y = f(∑_i w_i x_i + b)

dove x_i sono input, w_i sono pesi, b è il bias, e f è una funzione di attivazione (tipicamente la funzione gradino).

Il percettrone può apprendere funzioni linearmente separabili attraverso la regola di apprendimento:

w_i ← w_i + η(y_target - y_output)x_i

dove η è il tasso di apprendimento.

Minsky e Papert dimostrarono nel 1969 che un singolo percettrone non può calcolare XOR, limitando severamente il suo potere. Questo portò al "winter" dell'AI negli anni 1970.

Reti Multi-Strato e Backpropagation

I percettroni multi-strato (MLP) superano le limitazioni combinando strati di neuroni. Una MLP con un strato nascosto calcola:

y = f_2(W_2 f_1(W_1 x + b_1) + b_2)

dove f_1 e f_2 sono funzioni di attivazione non lineari.

Il teorema di approssimazione universale afferma che una MLP con un strato nascosto sufficientemente grande può approssimare qualunque funzione continua su un compatto arbitrariamente bene.

L'addestramento usa backpropagation per calcolare gradienti. Dato errore E = (1/2)||y - t||^2, la regola della catena dà:

∂E/∂w_ij = ∂E/∂y_j · ∂y_j/∂net_j · ∂net_j/∂w_ij

dove net_j = ∑_i w_ij x_i. Questo si propaga all'indietro attraverso gli strati:

δ_j^(l) = f'(net_j^(l)) ∑_k w_jk^(l+1) δ_k^(l+1)

Gli aggiornamenti dei pesi sono:

w_ij ← w_ij - η ∂E/∂w_ij = w_ij - η δ_j x_i

Funzioni di Attivazione

La scelta della funzione di attivazione è importante. Opzioni comuni:

Sigmoide: σ(x) = 1/(1+exp(-x)), range (0,1), derivata σ'(x) = σ(x)(1-σ(x))

Tanh: tanh(x) = (exp(x)-exp(-x))/(exp(x)+exp(-x)), range (-1,1), derivata 1-tanh^2(x)

ReLU: max(0,x), derivata 1 se x>0, 0 altrimenti

Leaky ReLU: max(αx,x) con α piccolo (es. 0,01)

Softmax (per output multiclasse): y_i = exp(x_i)/∑_j exp(x_j)

ReLU è diventata dominante per reti profonde perché mitiga il problema del gradiente che svanisce.

Reti Convoluzionali

Le reti neurali convoluzionali (CNN) sono specializzate per dati con struttura a griglia come immagini. Usano strati convoluzionali:

y_ij^k = f(∑*(m,n) w_mn^k x*(i+m,j+n) + b^k)

dove w^k è un filtro (kernel) appreso.

Le proprietà chiave sono:

Connettività sparsa: ogni neurone si connette solo a una regione locale dell'input

Condivisione dei parametri: lo stesso filtro è applicato ovunque

Equivarianza traslazionale: traslare l'input trasla l'output

I layer di pooling riducono risoluzione spaziale:

$y_{ij} = \max_{(m,n \in R_{ij})} x_{mn}$ (max pooling)

Le CNN moderne per visione (ResNet, VGG, Inception) impilano molti strati convoluzionali e di pooling, apprendendo gerarchie di features da bordi semplici a oggetti complessi.

Reti Ricorrenti e Memoria

Le reti ricorrenti (RNN) processano sequenze mantenendo uno stato nascosto:

$h_t = f(W_{hh} h_{(t-1)} + W_{xh} x_t + b_h)$ $y_t = W_{hy} h_t + b_y$

Le RNN semplici soffrono del problema del gradiente che svanisce/esplode per dipendenze a lungo termine. Le LSTM e GRU mitigano questo con architetture più sofisticate discusse precedentemente.

Le reti Hopfield sono reti ricorrenti simmetriche usate come memorie associative. Con pesi simmetrici $W_{ij} = W_{ji}$ e assenza di auto-connessioni $W_{ii} = 0$, aggiornamenti asincroni:

$s_i \leftarrow \text{sign}(\sum_j W_{ij} s_j)$

convergono a punti fissi che minimizzano l'energia:

$E = -(1/2) \sum_{(i,j)} W_{ij} s_i s_j$

Pattern memorizzati corrispondono a minimi locali. La capacità è circa 0,14N pattern per una rete di N neuroni.

Reti Generative e Autoencoders

Gli autoencoders apprendono rappresentazioni compresse dei dati. Consistono di un encoder che mappa input x a rappresentazione z:

$z = f_{enc}(x)$

e un decoder che ricostruisce:

$\hat{x} = f_{dec}(z)$

L'addestramento minimizza l'errore di ricostruzione:

$L = ||x - \hat{x}||^2$

I variational autoencoders (VAE) sono modelli generativi probabilistici. L'encoder produce parametri di una distribuzione:

q(z|x) = N(μ(x), Σ(x))

Il decoder definisce:

p(x|z) = N(f_dec(z), σ^2I)

L'obiettivo è massimizzare l'ELBO:

L = E_q[log p(x|z)] - D_KL(q(z|x)||p(z))

dove p(z) = N(0,I) è il prior.

I VAE possono generare nuovi campioni campionando z ~ p(z) e decodificando.

Le Generative Adversarial Networks (GAN) usano due reti in competizione: un generatore G che crea campioni falsi e un discriminatore D che distingue reali da falsi.

L'obiettivo minimax è:

min_G max_D E_x[log D(x)] + E_z[log(1-D(G(z)))]

Le GAN hanno prodotto immagini sintetiche di qualità impressionante ma sono notoriamente difficili da addestrare.

Meccanismi di Attenzione

Il meccanismo di attenzione permette alle reti di focalizzarsi su parti rilevanti dell'input. Nella forma più semplice:

α_i = exp(e_i) / ∑_j exp(e_j) c = ∑_i α_i h_i

dove h_i sono stati nascosti, e_i = score(h_i, s) misura rilevanza rispetto a query s, e c è il contesto atteso.

La funzione di scoring può essere:

Additiva: e_i = v^T tanh(W[h_i; s]) Moltiplicativa: e_i = h_i^T W s Dot product scalato: e_i = h_i^T s / √d

L'attenzione è centrale nei Transformers, permettendo modellare dipendenze a lungo raggio direttamente.

Conclusione Integrativa

La matematica della cognizione è un campo vasto e in rapida evoluzione che integra strumenti da numerose discipline. La teoria dell'informazione quantifica il contenuto informativo e i limiti di trasmissione. La probabilità bayesiana modella l'inferenza ottimale

sotto incertezza. La geometria caratterizza la struttura di spazi rappresentazionali. L'algebra lineare fornisce il linguaggio per rappresentazioni distribuite. I sistemi dinamici catturano la temporalità dei processi mentali. La teoria della decisione formalizza la scelta razionale. Le reti neurali artificiali forniscono modelli implementabili dei computazioni cognitive.

Questi framework matematici non sono meramente descrittivi ma forniscono teoria normativa: specificano cosa sarebbe ottimale dati vincoli e obiettivi. Confrontare comportamento umano con predizioni normative rivela dove gli umani sono ottimali, dove deviano sistematicamente, e quali vincoli computazionali o evolutivi potrebbero spiegare le deviazioni.

La matematizzazione ha trasformato lo studio della cognizione da un'impresa principalmente qualitativa a una scienza quantitativa e predittiva. I modelli matematici permettono predizioni precise, confronti quantitativi con dati, e accumulo cumulativo di conoscenza. Hanno anche guidato sviluppi tecnologici in intelligenza artificiale, interfacce cervello-computer, e sistemi di supporto decisionale.

Tuttavia, rimangono sfide profonde. Molti aspetti della cognizione, particolarmente quelli che coinvolgono significato, coscienza, e creatività, resistono ancora a caratterizzazione matematica soddisfacente. La relazione tra modelli matematici astratti e meccanismi neurali concreti è spesso poco chiara. Il trade-off tra fedeltà descrittiva e trattabilità matematica è perpetuo.

Ciononostante, il progresso è stato notevole. I modelli bayesiani hanno illuminato percezione, apprendimento e ragionamento. Le reti neurali artificiali hanno raggiunto performance umane o sovraumane su numerosi compiti. La teoria dei sistemi dinamici ha rivelato principi di auto-organizzazione nella cognizione. La teoria dell'informazione ha quantificato efficienza di codifica neurale. Questi successi suggeriscono che, pur con le sue limitazioni, la matematica continuerà a essere uno strumento indispensabile per comprendere la mente.

Sviluppi ulteriori

Sezione 1: Aspetti Emergenti e Contemporanei

Teoria delle Categorie e Composizionalità Concettuale

La teoria delle categorie fornisce un framework astratto per descrivere strutture matematiche e le relazioni tra di esse. Recentemente, è emerso interesse nell'applicare questo linguaggio alla cognizione, particolarmente per modellare la composizionalità concettuale.

Una **categoria** C consiste di:

- Una collezione di oggetti Ob(C)
- Per ogni coppia di oggetti A, B, un insieme di morfismi Hom(A,B)

- Un'operazione di composizione ◦ che soddisfa associatività
- Per ogni oggetto A, un morfismo identità id_A

Nel contesto cognitivo, gli oggetti possono rappresentare concetti o spazi rappresentazionali, mentre i morfismi rappresentano trasformazioni o mappature tra di essi.

Un **funtore** F: C → D mappa oggetti e morfismi da una categoria all'altra preservando la struttura:

- F(id_A) = id_F(A)
- F(g ◦ f) = F(g) ◦ F(f)

I funtori possono modellare come rappresentazioni in un dominio si mappano sistematicamente su un altro. Per esempio, il mapping da rappresentazioni linguistiche a rappresentazioni semantiche può essere modellato come funtore.

Le **trasformazioni naturali** catturano relazioni sistematiche tra funtori. Data due funtori F,G: C → D, una trasformazione naturale η: F → G assegna a ogni oggetto A in C un morfismo η_A: F(A) → G(A) tale che per ogni morfismo f: A → B, il diagramma commuta:

```
F(A) --η_A--> G(A)
 |             |
F(f)          G(f)
 |             |
 v             v
F(B) --η_B--> G(B)
```

Questo formalizza l'idea che certi mapping tra rappresentazioni sono "uniformi" attraverso diversi contenuti.

Composizionalità categoriale: Nel linguaggio naturale, il significato di espressioni complesse deriva composizionalmente dai significati delle parti. Il framework categoriale permette di formalizzare questo elegantemente. Per esempio, nella grammatica categoriale, ogni espressione ha un tipo (oggetto) e regole di combinazione (morfismi).

I tipi base potrebbero essere N (nome) e S (frase). Tipi complessi si costruiscono con il "slash": S/N è una funzione che prende un nome e produce una frase (verbo transitivo). La composizione funzionale corrisponde all'applicazione di morfismi.

Prodotti tensoriali nella teoria delle categorie monoidal forniscono un modo di combinare oggetti. Nel contesto di rappresentazioni vettoriali, il prodotto tensoriale u ⊗ v di due vettori crea uno spazio di dimensionalità maggiore che preserva informazione su entrambi i componenti mentre permettendo interazioni.

Per vettori u ∈ ℝ^m e v ∈ ℝ^n, il prodotto tensoriale u ⊗ v ∈ ℝ^(m×n) ha componenti: (u ⊗ v)_ij = u_i · v_j

Questo è stato proposto per modellare binding di ruoli nella rappresentazione di strutture (per esempio, legare "agente" e "Giovanni" in "Giovanni colpisce Maria").

Geometria dell'Informazione

La geometria dell'informazione studia famiglie di distribuzioni di probabilità come varietà differenziabili, fornendo una struttura geometrica intrinseca allo spazio statistico.

Consideriamo una famiglia parametrica di distribuzioni $\{p(x|\theta) : \theta \in \Theta\}$. Lo spazio dei parametri Θ può essere dotato di una struttura Riemanniana dove la metrica è la **matrice di informazione di Fisher**:

$$g_{ij}(\theta) = E_\theta[(\partial \log p(x|\theta)/\partial\theta_i)(\partial \log p(x|\theta)/\partial\theta_j)]$$

Questa metrica quantifica quanto rapidamente la distribuzione cambia quando i parametri variano. Per distribuzioni nella famiglia esponenziale:

$$p(x|\theta) = h(x) \exp(\theta^T T(x) - A(\theta))$$

la matrice di Fisher è:

$$g_{ij}(\theta) = \partial^2 A(\theta)/\partial\theta_i\partial\theta_j$$

dove $A(\theta)$ è la funzione di partizione logaritmica.

La **divergenza di Kullback-Leibler** può essere vista geometricamente. Mentre non è una distanza metrica (è asimmetrica), la sua approssimazione al secondo ordine è:

$$D_{KL}(p_\theta \,||\, p_{\theta+\delta\theta}) \approx (1/2) \sum_{ij} g_{ij}(\theta)\, \delta\theta_i\, \delta\theta_j$$

mostrando che la metrica di Fisher è la Hessiana locale della divergenza KL.

Connessioni duali: La geometria dell'informazione introduce il concetto di connessioni duali. La varietà statistica ha due connessioni affini naturali: la connessione esponenziale (e-connessione) e la connessione mista (m-connessione), che sono duali rispetto alla metrica di Fisher. Questa dualità è centrale nella teoria.

Per l'inferenza bayesiana, il **prior**, il **posterior** e il **predictive distribution** possono essere visti come punti su varietà statistiche. Il processo di aggiornamento bayesiano corrisponde a una proiezione geometrica.

Applicazioni cognitive: La geometria dell'informazione fornisce metriche naturali per quantificare cambiamenti nelle credenze durante l'apprendimento. La distanza di Fisher tra distribuzioni può misurare quanto un organismo ha "appreso". Il gradiente naturale, definito usando la metrica di Fisher, fornisce direzioni di apprendimento che sono invarianti rispetto a riparametrizzazioni:

$$\theta^{(t+1)} = \theta^{(t)} + \eta\, G^{(-1)}(\theta^{(t)})\, \nabla L(\theta^{(t)})$$

dove G è la matrice di Fisher e L è la loss function.

Principio dell'Energia Libera

Il **principio dell'energia libera** (Free Energy Principle, FEP), sviluppato da Karl Friston, propone che sistemi biologici minimizzano una quantità chiamata energia libera variazionale come principio unificante per percezione, azione e apprendimento.

Dato uno stato interno s, uno stato esterno ψ, e osservazioni sensoriali o generate da ψ, l'energia libera variazionale è definita come:

$F(s,o) = -E_q(ψ|s)[\log p(o,ψ)] + E_q(ψ|s)[\log q(ψ|s)]$

dove $q(ψ|s)$ è una distribuzione di riconoscimento approssimata (il modello interno dell'organismo) e $p(o,ψ)$ è il modello generativo.

L'energia libera può essere riscritta come:

$F(s,o) = -\log p(o|s) + D_{KL}(q(ψ|s) \,||\, p(ψ|o,s))$

Questo mostra che F è un limite superiore sulla surprisal negativa $-\log p(o)$ (l'energia libera della teoria dell'informazione). Minimizzare F quindi minimizza la surprisal e la divergenza KL tra il modello interno e il vero posterior.

La minimizzazione dell'energia libera si decompone in:

1. **Percezione**: Aggiornare gli stati interni s per minimizzare F dato o (inferenza percettiva)
2. **Azione**: Campionare azioni per cambiare o in modo da minimizzare F previsto (inferenza attiva)
3. **Apprendimento**: Aggiornare i parametri del modello generativo per minimizzare F nel tempo

Predizione ed errore di predizione: Il FEP implica che il cervello implementa una gerarchia di modelli generativi dove ogni livello predice l'attività del livello inferiore. Gli errori di predizione (la differenza tra predizione e input effettivo) propagano verso l'alto, mentre le predizioni propagano verso il basso.

Matematicamente, per una gerarchia di stati latenti $x^{(1)}, x^{(2)}, ..., x^{(n)}$, l'energia libera può essere espressa come somma di errori di predizione pesati per la precisione:

$F \approx \sum_i (1/2)\, ε_i^T Π_i ε_i$

dove $ε_i$ è l'errore di predizione al livello i e $Π_i$ è la matrice di precisione (inversa della covarianza del rumore).

Il cervello minimizza F aggiustando sia le predizioni (percezione) che campionando azioni per rendere gli input conformi alle predizioni (inferenza attiva).

Inferenza Attiva

L'**inferenza attiva** estende il FEP esplicitamente al dominio motorio. Invece di semplicemente inferire cause nascoste delle osservazioni, gli organismi agiscono per portare osservazioni in linea con le loro predizioni.

Questo risolve il problema classico del controllo motorio invertendo la direzione causale: invece di computare comandi motori da stati desiderati, il cervello predice stati sensoriali desiderati e i riflessi implementano automaticamente le azioni necessarie (riducendo l'errore di predizione propriocettivo).

Formalmente, data una policy (sequenza di azioni) π, l'organismo valuta l'energia libera attesa:

$G(\pi) = E_q[F(s,o) \mid \pi]$

e seleziona la policy che minimizza G. Questo può essere decomposto come:

$G(\pi) \approx -E_q[\log p(o|\pi)] + D_{KL}(q(\psi|\pi) \,||\, p(\psi))$

Il primo termine cattura l'utilità pragmatica (quanto le osservazioni attese sono preferite), mentre il secondo cattura la complessità epistematica (quanto le credenze cambiano).

Curiosità come information gain: Nel framework dell'inferenza attiva, comportamenti esplorativi emergono naturalmente come minimizzazione dell'energia libera attesa. La riduzione prevista dell'incertezza sugli stati nascosti (information gain) contribuisce a minimizzare G, spiegando la curiosità intrinseca.

L'information gain atteso è:

$IG(\pi) = E_q[D_{KL}(q(\psi|o,\pi) \,||\, q(\psi|\pi))]$

che misura quanto ci aspettiamo di apprendere eseguendo la policy π.

Quantum Cognition

Sebbene controverso, il framework della **quantum cognition** propone che alcune strutture matematiche della meccanica quantistica possano modellare fenomeni cognitivi che violano la probabilità classica.

Spazi di Hilbert: Invece di rappresentare stati cognitivi come distribuzioni di probabilità classiche su uno spazio di eventi, la quantum cognition li rappresenta come vettori in uno spazio di Hilbert H. Lo stato cognitivo è $|\psi\rangle \in H$.

Osservabili come operatori: Le domande o compiti corrispondono a operatori hermitiani (osservabili) A con autovalori λ_i e autovettori $|e_i\rangle$. La probabilità di ottenere risultato λ_i quando il sistema è nello stato $|\psi\rangle$ è:

$P(\lambda_i) = |\langle e_i|\psi\rangle|^2$

Principio di incompatibilità: Osservabili non commutanti ($AB \neq BA$) non possono essere misurati simultaneamente con precisione arbitraria. Questo è stato usato per modellare effetti di ordine nelle domande: rispondere alla domanda A prima della domanda B può dare risultati diversi che rispondere B prima di A.

Interferenza quantistica: In probabilità classica, la probabilità totale è la somma delle probabilità delle vie alternative:

$P(x) = \sum_i P(x|via_i)P(via_i)$

Nella probabilità quantistica, le amplificate di probabilità si sommano:

$\psi(x) = \sum_i \psi(x|via_i)\psi(via_i)$

e la probabilità è il modulo quadrato:

$P(x) = |\psi(x)|^2 = |\sum_i \psi(x|via_i)\psi(via_i)|^2$

Espandendo:

$P(x) = \sum i\, |\psi(x|via_i)|^2|\psi(via_i)|^2 + 2\sum\{i<j\} Re[\psi^*(x|via_i)\psi(x|via_j)\psi^*(via_i)\psi(via_j)]$

Il secondo termine è l'interferenza, assente nella probabilità classica. Questo è stato usato per modellare violazioni della legge di probabilità totale osservate in giudizi e decisioni umane.

Esempio: Effetto congiunzione: Il famoso problema di Linda (Tversky & Kahneman) mostra che le persone giudicano P(Linda è cassiera di banca e femminista) > P(Linda è cassiera di banca), violando logica probabilistica. Modelli quantistici possono catturare questo tramite interferenza:

$P(B \wedge F) = P(B)P(F|B) + \text{interferenza}$

dove l'interferenza positiva può far sì che $P(B \wedge F) > P(B)$.

Entanglement: Stati entangled sono stati congiunti che non possono essere fattorizzati. Per due sistemi A e B, uno stato separabile è $|\psi\rangle = |\psi_A\rangle \otimes |\psi_B\rangle$. Uno stato entangled non ha questa forma.

È stato proposto che certi processi cognitivi coinvolgono stati entangled dove, per esempio, la rappresentazione di un concetto non può essere separata dal contesto.

Cautela interpretativa: È importante notare che quantum cognition usa la matematica quantistica come formalismo, non implica necessariamente che il cervello opera con fenomeni quantistici fisici (superposizione, entanglement fisico a livello neuronale). Il dibattito continua su se questo è semplicemente un formalismo utile o indica qualcosa di più profondo.

Sezione 2: Neuroscienze Computazionali

Modelli Biofisici di Neuroni

La comprensione quantitativa dei neuroni richiede modelli che catturano le proprietà elettriche della membrana cellulare e la dinamica dei canali ionici.

Modello di Hodgkin-Huxley

Il modello di **Hodgkin-Huxley** (1952) descrive la generazione del potenziale d'azione nello squid giant axon attraverso equazioni differenziali che modellano le correnti ioniche.

La corrente totale attraverso la membrana è:

$I = C_m \, dV/dt + I_{ion}$

dove V è il potenziale di membrana, C_m è la capacitanza della membrana (tipicamente ~1 $\mu F/cm^2$), e I_{ion} è la somma delle correnti ioniche.

Le correnti ioniche principali sono sodio (Na^+), potassio (K^+), e leak:

$I_{ion} = g_{Na} m^3 h(V - E_{Na}) + g_K n^4(V - E_K) + g_L(V - E_L)$

dove:

- g_{Na}, g_K, g_L sono le conduttanze massime
- E_{Na}, E_K, E_L sono i potenziali di inversione
- m, h, n sono variabili di gating che rappresentano la frazione di canali aperti

Le variabili di gating evolvono secondo:

$dx/dt = \alpha_x(V)(1-x) - \beta_x(V)x$

dove $x \in \{m,h,n\}$ e α_x, β_x sono funzioni del voltaggio che descrivono le rate di apertura e chiusura.

Per esempio, per la variabile m (attivazione del sodio):

$\alpha_m(V) = 0{,}1(V+40)/(1 - \exp(-(V+40)/10))$ $\beta_m(V) = 4 \exp(-(V+65)/18)$

Simulazione del potenziale d'azione: Con questi parametri, il modello riproduce accuratamente la forma, ampiezza (~100 mV), e durata (~1 ms) del potenziale d'azione, oltre a proprietà come periodo refrattario e soglia.

Modello Integrate-and-Fire

Per simulazioni di grandi reti, il modello di Hodgkin-Huxley è computazionalmente costoso. Il modello **leaky integrate-and-fire** (LIF) fornisce un'astrazione più semplice:

$C_m \, dV/dt = -g_L(V - E_L) + I_{ext}$

dove I_{ext} è la corrente sinaptica o iniettata esternamente. Quando V raggiunge una soglia V_{th}, il neurone emette uno spike e V viene resettato a V_{reset}.

In forma adimensionale, con $\tau_m = C_m/g_L$ (costante di tempo della membrana, tipicamente 10-20 ms):

$\tau_m \, dV/dt = -(V - E_L) + R_m I_{ext}$

Quando V = V_th, allora V → V_reset.

Varianti: Il modello LIF può essere esteso per catturare più fenomeni:

1. **Exponential integrate-and-fire (EIF)**: Aggiunge un termine esponenziale per catturare l'upstroke rapido dello spike:

   $C_m \, dV/dt = -g_L(V - E_L) + g_L \Delta_T \exp((V - V_T)/\Delta_T) + I_{ext}$

   dove Δ_T è lo sharpness dello spike e V_T è la soglia effettiva.

2. **Adaptive integrate-and-fire**: Aggiunge una corrente di adattamento w che aumenta con ogni spike:

   $\tau_m \, dV/dt = -(V - E_L) + I_{ext} - w \; \tau_w \, dw/dt = a(V - E_L) - w$

   Quando V = V_th: V → V_reset, w → w + b

   Questo cattura spike frequency adaptation.

Modello di Izhikevich

Il modello di **Izhikevich** (2003) bilancia realismo biologico ed efficienza computazionale:

$dV/dt = 0{,}04V^2 + 5V + 140 - u + I \quad du/dt = a(bV - u)$

Quando $V \geq 30$ mV: V → c, u → u + d

I quattro parametri (a,b,c,d) possono essere aggiustati per riprodurre una vasta gamma di comportamenti neuronali osservati sperimentalmente: regular spiking, intrinsically bursting, fast spiking, chattering, etc.

Per esempio:

- Regular spiking (corteccia): a=0,02, b=0,2, c=-65, d=8
- Fast spiking (interneuroni): a=0,1, b=0,2, c=-65, d=2
- Intrinsically bursting: a=0,02, b=0,2, c=-55, d=4

Sinapsi e Plasticità

Le connessioni tra neuroni (sinapsi) sono dinamiche e modificabili. Modelli matematici catturano come la forza sinaptica cambia con l'attività.

Spike-Timing-Dependent Plasticity (STDP)

La **STDP** è una forma di plasticità hebbiana dove il cambiamento nella forza sinaptica dipende dal timing relativo degli spike pre- e post-sinaptici.

Se il neurone pre-sinaptico spara $\Delta t = t_{post} - t_{pre}$ prima del post-sinaptico ($\Delta t > 0$, causale), la sinapsi viene potenziata. Se spara dopo ($\Delta t < 0$, anticausale), viene depressa.

La finestra temporale di STDP è tipicamente modellata come:

$\Delta w = A_+ \exp(-\Delta t/\tau_+)$ se $\Delta t > 0$ (LTP) $\Delta w = -A_- \exp(\Delta t/\tau_-)$ se $\Delta t < 0$ (LTD)

dove A_+, A_- controllano l'ampiezza e τ_+, τ_- controllano la larghezza della finestra (tipicamente $\tau_+ \approx \tau_- \approx 20$ ms).

Apprendimento di sequenze: STDP può apprendere sequenze temporali: se neuroni sparano ripetutamente in sequenza A→B→C, le connessioni A→B e B→C vengono rafforzate mentre B→A e C→B vengono indebolite, creando una "chain" direzionale.

Plasticità Omeostatica

Per evitare instabilità (rinforzo illimitato o depressione completa), meccanismi di **plasticità omeostatica** regolano la forza sinaptica verso un set-point:

$\tau_h \, d\theta/dt = (r - r_{target})$

dove θ è la soglia neuronale, r è il firing rate, e r_{target} è il target rate. Questo implementa un controllo a feedback negativo che stabilizza l'attività.

Oscillazioni Neurali e Sincronizzazione

Le oscillazioni sono ubique nell'attività cerebrale, spaziando da ritmi delta (1-4 Hz) nel sonno profondo a oscillazioni gamma (30-100 Hz) nella cognizione.

Modello di Wang-Buzsáki

Per interneuroni che generano oscillazioni gamma, il modello di Wang-Buzsáki include conduttanze sinaptiche GABAergiche e correnti ioniche specifiche:

$C_m \, dV/dt = -I_{Na} - I_K - I_L - I_{syn}$

dove $I_{syn} = g_{syn} \, s(V - E_{syn})$ è la corrente sinaptica e s è la frazione di recettori aperti che evolve secondo:

$ds/dt = \alpha(1-s)\sum_j \delta(t - t_j) - s/\tau_s$

dove la somma è sugli spike ricevuti e $\tau_s \approx 5$-10 ms.

In una rete di interneuroni accoppiati, le oscillazioni gamma emergono dall'interazione tra eccitazione ricorrente e inibizione.

Sincronizzazione di Fase

Due oscillatori sono sincronizzati in fase se la differenza di fase rimane limitata. Per oscillatori accoppiati con fasi θ_1, θ_2:

$d\theta_1/dt = \omega_1 + K \sin(\theta_2 - \theta_1)$ $d\theta_2/dt = \omega_2 + K \sin(\theta_1 - \theta_2)$

La differenza di fase $\Delta\theta = \theta_2 - \theta_1$ evolve come:

$d(\Delta\theta)/dt = \Delta\omega - 2K \sin(\Delta\theta)$

dove $\Delta\omega = \omega_2 - \omega_1$. Per $K > |\Delta\omega|/2$, esiste un punto fisso stabile con $\Delta\theta$ costante (sincronizzazione).

Coerenza di fase: Per quantificare la sincronizzazione in dati sperimentali, si usa la misura di coerenza di fase (Phase Locking Value):

$PLV = |1/N \sum_n \exp(i(\varphi_1(t_n) - \varphi_2(t_n)))|$

dove φ_1, φ_2 sono le fasi istantanee di due segnali. PLV varia tra 0 (nessun locking) e 1 (locking perfetto).

Connettività Funzionale vs Strutturale

La **connettività strutturale** descrive le connessioni anatomiche (assoni, dendriti) tra regioni cerebrali, mentre la **connettività funzionale** descrive correlazioni statistiche nell'attività.

Matrice di Connettività

La connettività può essere rappresentata come grafo pesato con matrice di adiacenza W dove W_{ij} rappresenta la forza della connessione da i a j.

Per connettività strutturale, W_{ij} potrebbe essere il numero di fibre (tractografia DTI) o la densità sinaptica.

Per connettività funzionale, W_{ij} è tipicamente la correlazione:

$W_{ij} = corr(x_i(t), x_j(t))$

dove $x_i(t)$ è il segnale (BOLD fMRI, LFP) dalla regione i.

Modelli di Connettività Dinamica Causale (DCM)

La DCM modella come regioni cerebrali influenzano l'una l'altra nel tempo. Il modello neuronale più semplice è lineare:

$dz/dt = Az + Cu$

dove z è il vettore di stati neurali, A è la matrice di connettività intrinseca, C specifica come input esterni u influenzano il sistema.

Le modulazioni possono essere introdotte:

$dz/dt = (A + \sum_j u_j B^{(j)})z + Cu$

dove $B^{(j)}$ descrive come l'input u_j modula la connettività.

Il modello neuronale è legato ai dati osservati (BOLD fMRI) tramite una funzione di osservazione non lineare (modello emodinamico di Balloon-Windkessel).

L'inferenza bayesiana stima i parametri (A, B, C) dai dati, quantificando l'evidenza per diversi modelli di connettività causale.

Codifica e Decodifica Neurale

Curve di Tuning

Neuroni in molte aree sensoriali e motorie hanno **curve di tuning**: la loro risposta dipende da una variabile di stimulus s. La funzione di tuning f(s) descrive il firing rate medio:

$r = f(s) + \varepsilon$

dove ε è rumore (tipicamente Poisson: $var(\varepsilon) \propto r$).

Per esempio, neuroni nella corteccia visiva primaria hanno tuning per orientamento:

$f(\theta) = r_{max} \exp(-(\theta - \theta_{pref})^2/(2\sigma^2))$

dove θ_{pref} è l'orientamento preferito e σ determina la larghezza del tuning.

Codifica di Popolazioni

Una popolazione di neuroni con curve di tuning sovrapposte può rappresentare stimoli accuratamente. Data una popolazione di N neuroni con firing rates r_i, lo stimulus può essere decodificato come:

$\hat{s} = \sum_i w_i r_i / \sum_i r_i$

dove $w_i = s_i$ è lo stimulus preferito del neurone i (population vector decoding).

Per decodifica bayesiana ottimale:

$\hat{s}_{MAP} = \text{argmax}_s\, p(s|r) \propto \text{argmax}_s\, p(r|s)p(s)$

Per neuroni Poisson indipendenti:

$p(r|s) = \prod_i (f_i(s)^{r_i} \exp(-f_i(s)))/r_i!$

quindi:

$\log p(r|s) = \sum_i [r_i \log f_i(s) - f_i(s)]$

Informazione di Fisher della Popolazione

L'**informazione di Fisher** quantifica quanta informazione una popolazione porta su uno stimulus. Per una popolazione di neuroni Poisson:

$I(s) = \sum_i (f_i'(s))^2/f_i(s)$

dove f_i'(s) = df_i/ds. L'informazione di Fisher è inversamente proporzionale alla varianza del miglior stimatore non distorto (Cramér-Rao bound):

var(ŝ) ≥ 1/I(s)

Per curve di tuning gaussiane con ampiezza r_max e larghezza σ distribuite uniformemente in orientamento:

I(θ) ≈ N r_max/(2σ²)

mostrando che l'informazione scala linearmente con la dimensione della popolazione N e inversamente con la larghezza del tuning σ.

Reti Neurali Ricorrenti e Attrattori

Le reti neurali ricorrenti possono implementare memoria e computazione attraverso dinamiche attrattoriali.

Rete di Hopfield Continua

La rete di Hopfield continua evolve secondo:

τ dV_i/dt = -V_i + ∑_j W_ij σ(V_j) + I_i

dove V_i è il potenziale del neurone i, σ è una funzione sigmoidale, W_ij sono i pesi sinaptici, e I_i è input esterno.

Per pesi simmetrici W_ij = W_ji, la rete minimizza la funzione energia:

E = -(1/2)∑_{ij} W_ij σ(V_i)σ(V_j) - ∑_i I_i σ(V_i) + (1/τ)∑_i ∫_0^{V_i} σ^(-1)(v)dv

Gli stati stabili corrispondono a minimi locali di E.

Attrattori Continui per Memoria di Lavoro

Per implementare memoria di lavoro di variabili continue (per esempio, posizione spaziale), si usano attrattori continui ("line attractors" o "ring attractors").

In una rete con connettività a Mexican hat:

W(x,x') = A_0 exp(-(x-x')²/(2σ_E²)) - A_1 exp(-(x-x')²/(2σ_I²))

con σ_E < σ_I (eccitazione più stretta dell'inibizione), la rete sviluppa un continuum di stati stabili (un "bump" di attività può essere sostenuto a qualsiasi posizione x).

Questo è stato proposto come meccanismo per memoria di lavoro spaziale nell'ocular motor system e nella corteccia prefrontale.

Apprendimento Profondo e Neuroscienze

L'apprendimento profondo e le neuroscienze si informano reciprocamente, sebbene le reti artificiali siano semplificazioni drastiche dei circuiti biologici.

Confronti tra Reti Artificiali e Biologiche

Similarità:

- Rappresentazioni gerarchiche (features semplici → complesse)
- Plasticità sinaptica (backpropagation vs STDP)
- Rappresentazioni distribuite

Differenze:

- Backpropagation richiede gradienti esatti propagati all'indietro; il cervello probabilmente usa approssimazioni locali
- Neuroni biologici hanno spike temporalmente precisi; neuroni artificiali usano firing rates
- Il cervello integra feedback top-down estensivo; molte reti feed-forward
- Apprendimento biologico è tipicamente online e continuo; training artificiale spesso batch e episodico

Feedback Alignment e Plausibilità Biologica

Backpropagation standard richiede che i pesi backward siano trasposte dei pesi forward:

$$\delta^l = (W^{l+1})^T \delta^{l+1} \odot f'(z^l)$$

Feedback alignment (Lillicrap et al., 2016) mostra che pesi backward casuali B^{l+1} (non legati a W^{l+1}) possono comunque permettere apprendimento:

$$\delta^l = B^{l+1} \delta^{l+1} \odot f'(z^l)$$

Sorprendentemente, i pesi forward si allineano gradualmente con i pesi backward durante training, permettendo aggiornamenti efficaci. Questo è più biologicamente plausibile poiché non richiede simmetria di pesi.

Predictive Coding

Un framework alternativo più biologicamente plausibile è **predictive coding**. In una gerarchia, ogni livello predice l'attività del livello inferiore:

$$\hat{y}^l = f(W^l x^{l+1})$$

Gli errori di predizione:

$$\varepsilon^l = y^l - \hat{y}^l$$

propagano verso l'alto (feedforward) mentre le predizioni propagano verso il basso (feedback).

L'apprendimento minimizza:

$E = \sum_l (1/2)||\varepsilon^l||^2$

Le regole di aggiornamento:

$\Delta x^{l+1} \propto (W^l)^T \varepsilon^l - \varepsilon^{l+1} \Delta W^l \propto \varepsilon^l (x^{l+1})^T$

sono locali e potrebbero essere implementate da circuiti corticali con connessioni feedforward (errori) e feedback (predizioni).

Sezione 3: Cognizione Incarnata e Situata

Sistemi Dinamici della Percezione-Azione

La cognizione incarnata enfatizza che la cognizione emerge dall'interazione continua tra corpo, cervello e ambiente. I sistemi dinamici forniscono il linguaggio matematico per modellare questi accoppiamenti.

Coordinazione Sensori-Motoria

Il **coordinamento** tra componenti emerge da accoppiamenti dinamici. Un esempio classico è la coordinazione bimanuale studiata da Kelso. Quando le persone oscillano due dita, solo due pattern stabili emergono: in-fase ($\phi = 0$) e anti-fase ($\phi = \pi$).

Il modello HKB (Haken-Kelso-Bunz) cattura questo con l'equazione:

$d\phi/dt = -a \sin(\phi) - 2b \sin(2\phi)$

dove ϕ è la differenza di fase relativa tra le dita.

I punti fissi ($d\phi/dt = 0$) sono soluzioni di:

$a \sin(\phi) + 2b \sin(2\phi) = 0$

Per $b > 0$, ci sono sempre soluzioni a $\phi = 0, \pi$ (in-fase e anti-fase). La stabilità è determinata da:

$d\phi/dt|_{\phi^*} = -a \cos(\phi^*) - 4b \cos(2\phi^*)$

Per $b/a < 1/4$, entrambi i pattern sono stabili. Per $b/a > 1/4$, solo in-fase è stabile.

Quando la frequenza di oscillazione aumenta (parametro di controllo), il sistema subisce una **transizione di fase**: anti-fase diventa instabile e il sistema salta spontaneamente a in-fase. Questo è un esempio di biforcazione nel comportamento motorio.

Locomozione e CPG

I **central pattern generators** (CPG) sono circuiti neurali che producono output ritmici senza input ritmici. Sono centrali nella locomozione.

Un modello semplice di CPG è basato su oscillatori accoppiati. Per locomozione quadrupede, quattro oscillatori (una per zampa) con accoppiamenti:

$d\theta_i/dt = \omega + \sum_j K_{ij} \sin(\theta_j - \theta_i - \phi_{ij})$

dove θ_i è la fase dell'oscillatore i, ω è la frequenza intrinseca, K_{ij} è la forza di accoppiamento, e ϕ_{ij} è il phase lag desiderato.

Per il trotto, $\phi_{ij} = \pi$ tra zampe diagonali e $\phi_{ij} = 0$ tra zampe laterali. Per il galoppo, le zampe anteriori sono sincronizzate ($\phi = 0$) e precedono le posteriori ($\phi = \pi/2$).

La transizione tra andature (walk → trot → gallop) con l'aumentare della velocità può essere modellata come biforcazioni nel sistema dinamico quando la frequenza ω aumenta.

Teoria delle Affordances

Affordances (Gibson) sono opportunità di azione offerte dall'ambiente relative alle capacità dell'agente. La matematica dell'affordance cerca di formalizzare quando un'azione è possibile.

Rapporti Intrinseci

Le affordances dipendono da **rapporti intrinseci** tra dimensioni ambientali e dimensioni corporee. Per esempio, una sedia è "sedibile" se:

$0{,}8 < h_{sedile}/l_{gamba} < 1{,}2$

dove h_{sedile} è l'altezza della sedia e l_{gamba} è la lunghezza della gamba.

Una porta è "passabile" se:

$w_{porta}/w_{spalle} > 1{,}3$

dove w_{porta} è la larghezza della porta e w_{spalle} è la larghezza delle spalle (con margine di sicurezza).

Questi rapporti sono dimensionless e invarianti rispetto a scaling, catturando l'idea che affordances sono relazionali, non proprietà assolute.

Funzioni di Possibilità

Le affordances possono essere formalizzate come **funzioni di possibilità** $A(s,a,c)$ che mappano stato dell'ambiente s, azione a, e capacità dell'agente c a {0,1}:

$A(s,a,c) = 1$ se l'azione a è possibile nello stato s con capacità c $A(s,a,c) = 0$ altrimenti

Per esempio, per l'affordance "raggiungibile":

$A_{reach}(s,a,c) = 1$ se $dist(\text{agente}, \text{oggetto}) \leq \text{lunghezza_braccio}$ $A_{reach}(s,a,c) = 0$ altrimenti

Una versione probabilistica o fuzzy può catturare incertezza:

$A(s,a,c) \in [0,1]$

rappresentando il grado di affordance.

Percezione Ecologica e Tau

Nella **teoria ecologica della percezione**, la percezione è diretta dall'informazione ottica disponibile nell'ambiente, non da ricostruzioni interne.

Variabile Tau

La variabile **tau** (τ) quantifica il tempo al contatto durante approccio a un oggetto:

$\tau = z/(dz/dt)$

dove z è la distanza dall'oggetto. Se la velocità è costante, τ è letteralmente il tempo rimanente al contatto.

Criticamente, τ può essere percepito direttamente dalla rate di espansione dell'immagine retinica. Per un oggetto di dimensione fissa r, l'angolo visivo θ soddisfa:

$\tan(\theta/2) = r/(2z)$

Per angoli piccoli, $\theta \approx r/z$, quindi:

$d\theta/dt = -(r/z^2)(dz/dt) = -(1/z)(dz/dt)\theta = -\theta/\tau$

quindi:

$\tau = -\theta/(d\theta/dt)$

Questo mostra che τ è specificato dalla rate di cambiamento dell'angolo visivo relativa all'angolo stesso, una quantità direttamente disponibile nell'array ottico.

Strategia tau-coupling: Per azioni di intercettazione o frenata, organismi usano una strategia di mantenere τ costante o seguire una traiettoria specifica. Per frenata con tempo di arresto τ_{final}:

$\dot{\tau} = k(\tau - \tau_{final}/k)$

produce decelerazione graduale fino all'arresto. Evidenze empiriche supportano che giocatori di cricket, gabbiani che si tuffano, e guidatori usano strategie basate su tau.

Embodied Cognition e Schema Motori

Gli **schemi motori** sono pattern di azione generalizzati che possono essere parametrizzati per diversi contesti.

Generalized Motor Programs

Un generalized motor program (GMP) specifica la struttura temporale relativa di un'azione, mentre parametri specificano dettagli:

Action(t) = Amplitude · Pattern(Duration · t, Parameters)

Per esempio, firmare il proprio nome:

- Pattern: la forma relativa della firma
- Amplitude: quanto grande
- Duration: quanto veloce
- Parameters: inclinazione, etc.

La struttura invariante (Pattern) è preservata attraverso scaling spaziale e temporale.

Schema Theory di Schmidt

La schema theory propone che attraverso pratica, si sviluppano relazioni (schemi) tra:

1. **Condizioni iniziali** (posizione del corpo, proprietà oggetti)
2. **Parametri di risposta** (forza, velocità, durata)
3. **Outcome sensoriali** (dove l'oggetto è atterrato)

Il **recall schema** genera parametri appropriati per outcome desiderati:

Parameters = f_recall(InitialConditions, DesiredOutcome)

Il **recognition schema** genera aspettative sensoriali:

ExpectedFeedback = f_recog(InitialConditions, Parameters)

Durante esecuzione, il feedback effettivo è confrontato con le aspettative per rilevare errori.

Matematicamente, questi possono essere modellati come funzioni apprese attraverso regressione:

p̂ = W_recall [x_init, x_desired] ŷ = W_recog [x_init, p]

dove W_recall e W_recog sono matrici di peso apprese dall'esperienza.

Optimal Feedback Control

La teoria moderna del controllo motorio propone che il sistema nervoso risolve problemi di **optimal feedback control** (OFC).

Formulazione del Problema

Dato uno stato x(t) (posizione, velocità di arti), controlli u(t) (comandi motori), e obiettivo x_goal, si cerca la traiettoria ottimale che minimizza:

J = φ(x(T)) + ∫_0^T L(x(t),u(t)) dt

soggetto a dinamiche:

dx/dt = f(x,u)

Il termine finale φ(x(T)) penalizza deviazioni dall'obiettivo, mentre L(x,u) = x^T Q x + u^T R u penalizza deviazioni durante il movimento e sforzo di controllo.

Linear Quadratic Regulator (LQR)

Per dinamiche lineari e costi quadratici:

dx/dt = Ax + Bu J = ∫_0^∞ (x^T Q x + u^T R u) dt

la soluzione è feedback lineare:

u(t) = -Kx(t)

dove K = R^(-1) B^T P e P soddisfa l'equazione di Riccati:

A^T P + PA - PBR^(-1)B^T P + Q = 0

Questa è la base matematica per controllo ottimale.

Minimum Jerk Model

Il **minimum jerk model** (Flash & Hogan, 1985) propone che i movimenti punto-a-punto minimizzano il jerk (derivata terza della posizione):

$J = (1/2) \int_0^T (d^3x/dt^3)^2 \, dt$

soggetto a condizioni al contorno x(0), ẋ(0), x(T), ẋ(T).

Il calcolo variazionale dà la soluzione:

$x(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5$

un polinomio di quinto grado. Questo produce profili di velocità a campana ("bell-shaped") che corrispondono bene ai movimenti umani osservati.

Per movimenti 3D, si minimizza jerk indipendentemente in ogni dimensione, predicendo traiettorie relativamente dritte (principio di linearità) e velocità sincronizzate tra dimensioni (principio di isocronia).

Optimal Control con Incertezza (SOC)

Il **stochastic optimal control** (SOC) estende OFC considerando rumore sensoriale e motorio.

Signal-Dependent Noise

Il rumore nei comandi motori è proporzionale all'ampiezza del comando (Weber's law per il controllo motorio):

$u_{actual} = u_{commanded} + \varepsilon$

dove $\varepsilon \sim N(0, \sigma^2 u^2)$ con varianza che scala con u^2.

Questo implica che controlli grandi (movimenti rapidi, forti) sono più rumorosi. Il compromesso velocità-accuratezza (Fitts' law) emerge naturalmente: per colpire piccoli target, si devono usare controlli ridotti (più lenti) per mantenere precisione.

Minimum Variance Model

Per movimenti con noise dipendente dal segnale, il controllo ottimale minimizza la varianza della posizione finale:

$J = E[(x(T) - x_{goal})^2 + \int_0^T c\, u(t)^2\, dt]$

dove c controlla il costo dello sforzo.

La soluzione mostra che:

1. Correzioni feedback sono ottimali solo per deviazioni che influenzano il goal finale
2. Deviazioni ortogonali al goal sono tollerate (variabilità non controllata)

Questo spiega **minimum intervention principle**: il sistema corregge solo errori rilevanti per il task, non tutti gli errori.

Coordinazione Interpersonale

La cognizione incarnata si estende alla cognizione sociale, dove individui si coordinano attraverso accoppiamento dinamico.

Sincronizzazione di Oscillatori

Durante interazioni sociali (conversazione, musica), gli individui sincronizzano i loro movimenti e ritmi. Questo può essere modellato come oscillatori accoppiati:

$d\theta_1/dt = \omega_1 + K_{12} \sin(\theta_2 - \theta_1)$ $d\theta_2/dt = \omega_2 + K_{21} \sin(\theta_1 - \theta_2)$

La sincronizzazione emerge quando l'accoppiamento è sufficientemente forte.

Joint Action e Predizione

Nelle **joint actions**, i partner predicono le azioni dell'altro. Modelli bayesiani della cognizione sociale propongono che:

$P(goal_{other}|observation) \propto P(observation|goal_{other})\, P(goal_{other})$

Combinando predizioni sulle azioni del partner con i propri piani:

$u_{own} = \text{argmin}_u [\text{Cost}_{own}(u) + \lambda \text{Cost}_{coordination}(u, u_{predicted_other})]$

dove λ pesa l'importanza del coordinamento.

Morfologia Computazionale

Il corpo stesso può eseguire computazioni (**morphological computation**), riducendo il carico sul controllo neurale.

Passive Dynamic Walking

Robots passive dynamic walkers scendono pendii senza motori o controllo attivo, usando solo gravità e la loro morfologia meccanica. Le equazioni di movimento per un modello a gamba rigida sono:

$I_{leg} \ddot{\theta}_{stance} = mgl \sin(\theta_{stance})$ $I_{leg} \ddot{\varphi} = mgl \sin(\varphi) - mg(L+r \cos(\varphi))\sin(\theta_{stance})$

dove θ_{stance} è l'angolo della gamba di supporto, φ è l'angolo relativo della gamba oscillante.

Con parametri appropriati (distribuzione massa, lunghezza gamba), emergono cicli limite stabili che corrispondono a walking naturale. Questo dimostra che la morfologia del corpo contribuisce alla stabilità e pattern del movimento, non solo il controllo neurale.

Compliance e Impedenza

Articolazioni biologiche non sono rigide ma hanno **compliance** (cedevolezza) variabile. La relazione forza-posizione è:

$F = K(x - x_0) + B(\dot{x})$

dove K è stiffness (rigidità), B è damping (smorzamento), e x_0 è la posizione di equilibrio.

Il controllo di impedenza permette al sistema nervoso di regolare K e B, non solo la forza F. Questo è cruciale per interazione con ambienti incerti: alta compliance permette adattamento passivo a perturbazioni.

La co-contrazione di muscoli agonisti e antagonisti aumenta stiffness:

$K = K_0 + \alpha(u_{agonist} + u_{antagonist})$

dove u sono le attivazioni muscolari. Questo permette trade-off tra stabilità (alta K) e flessibilità (bassa K).

Sintesi e Prospettive

Queste tre sezioni integrative evidenziano la crescente sofisticazione e l'ampiezza della matematizzazione della cognizione.

Aspetti emergenti e contemporanei mostrano come nuovi formalismi matematici (teoria delle categorie, geometria dell'informazione, principio dell'energia libera) forniscono prospettive unificanti che connettono fenomeni precedentemente disparati.

Neuroscienze computazionali rivelano come modelli matematici a molteplici scale – da canali ionici singoli a reti estese – catturano la complessità del substrato neurale della cognizione, permettendo predizioni testabili e comprensione meccanicistica.

Cognizione incarnata sottolinea che la matematica della cognizione deve considerare l'intera loop percezione-cognizione-azione-ambiente, non solo processi "nella testa". I sistemi dinamici forniscono il linguaggio naturale per questa prospettiva integrata.

La frontiera futura richiederà:

1. **Integrazione cross-scala**: Collegare matematica molecolare, cellulare, circuitale, sistemica e comportamentale
2. **Modelli personalizzati**: Catturare variabilità individuale, non solo medie di popolazione
3. **Ecologia e validità**: Estendere modelli da laboratorio a comportamento naturalistico
4. **Cognizione collettiva**: Modellare matematicamente gruppi, cultura, società
5. **Coscienza**: Affrontare quantitativamente il "hard problem"

La matematica rimane lo strumento più potente per trasformare l'intuizione qualitativa sulla mente in scienza predittiva e cumulativa.